

Increasing interpretability in XAI

Addressing the design principles for interactive
XUIs to increase interpretability in XAI for end-users

Yining Ren

Increasing interpretability in XAI

Addressing the design principles for interactive
XUIs to increase interpretability in XAI for
end-users

by

Yining Ren

to obtain the degree of Master of Science in

Management of Technology,
Faculty of Technology, Policy and Management,

at the Delft University of Technology,
to be defended publicly on Friday May 17th, 2024 at 14:00 PM.

Student number: 5659507

Thesis committee: Dr. Y. (Aaron) Ding TU Delft, Chair & First Supervisor
Dr. J. Zatarain Salazar TU Delft, Second Supervisor
Dr. M. Westberg TU Delft, Advisor

Preface

This thesis marks the culmination of my journey exploring the topic of enhancing interpretability for end-users in Explainable Artificial Intelligence (XAI) through the use of an Explainable User Interface (XUI), leading to my Master's degree in Management of Technology. My decision to delve into this area stemmed from a desire to move away from traditional thesis frameworks and create something that not only addresses theoretical concerns but also has practical applications. The topic of (X)AI not only aligned with my personal interests and background but also with the master's program, and the findings from this study are intended to guide future developers in XAI and XUI design.

Throughout this research, I had the privilege of collaborating with the team at FOKUS. Their insights and feedback were crucial in shaping the practical aspects of the XUI design, ensuring that the theoretical proposals were grounded in real-world applications and that the solutions were both theoretically sound and practically viable. I extend my heartfelt thanks to the entire FOKUS team for their time and effort in this collaboration. I also want to thank all the participants who generously offered their time and expertise to participate in the interviews.

I owe immense gratitude to my advisor, Marcus Westberg, whose expert guidance and unwavering support were instrumental throughout the research and writing processes. I also extend my gratitude to my committee members, Aaron Ding, Jazmin Zatarain Salazar, and Rens Kortmann, for their invaluable support, feedback, and guidance throughout this journey. Their understanding and encouraging words significantly impacted my ability to persevere through all circumstances.

Furthermore, I want to thank my family and friends for their unconditional love and support during my academic years. Aunty Nuan, thank you for always believing in me, encouraging me, and providing nourishing meals over the years. To Mani, Shadin, and Rishab, the MOT journey would not have been the same without you, and I am so grateful to have experienced it alongside you. Similarly, to all my friends, thank you for all the conversations and the much-needed distractions in between. Lastly, I would like to thank Heiko for always being there for me and being my rock throughout everything.

I hope you enjoy reading this report!

*Yining Ren
Delft, May 2024*

Executive Summary

Background

Recent advancements in machine learning, particularly deep learning models, have significantly enhanced AI's capabilities, leading to more complex and less explainable algorithms. The rise of Explainable AI (XAI) seeks to address these challenges by making AI's decision-making processes transparent and interpretable. Two knowledge gaps were identified in the domain of XAI. The first knowledge gap presents itself in that the research in XAI has predominantly catered to the needs of those with a technical background, often overlooking the vast majority of users who lack technical expertise, the end-user. User interfaces are identified as crucial tools for enhancing the interpretability of AI explanations, and likewise, interactivity is also identified a pivotal factor in enhancing interpretability in XAI. The second gap shows that while user-centric design principles for eXplainable User Interfaces (XUIs) are theoretically well-recognized, their practical implementation in real-world settings where users interact with AI systems is still insufficiently explored. Hence, based on the knowledge gaps identified, the following main research questions was formed:

How can interactive design principles be strategically applied to Explainable User Interfaces (XUI) to enhance the interpretability of Explainable Artificial Intelligence (XAI) for end-users?

Research approach

The research adopted the Design Science Research Method (DSRM) as outlined by Peffers et al., 2007, guiding the development of the XUI for FOKUS. Initially, an extensive literature review was conducted, focusing on factors related to interpretability, types of interaction, and design principles in XAI. This was followed by defining the scope of the XUI through the exploration of various XAI methods and the conceptualisation of design elements tailored to the FOKUS use case, which involves detecting myocardial infarctions (MIs) using Electrocardiogram (ECG) data. Key questions about what, how, and to whom the information should be explained led to the selection of specific design principles proposed by Chromik and Butz, 2021: complementary naturalness, responsiveness through progressive disclosure, and flexibility in explanation methods, while sensitivity to context was excluded due to current system limitations. This methodological approach resulted in a theoretical framework that effectively bridges interpretability with interactivity, underpinned by selected design principles, creating a comprehensive roadmap for developing an XUI that meets both the technical requirements of XAI and the practical needs of end-users. The XUI's application in the FOKUS use case utilised SHAP, LIME, and Grad-CAM to demonstrate flexibility in explanation methods, ensuring a diverse yet non-overwhelming analysis. It integrated complementary naturalness by combining textual and visual explanations, flexibility through multiple ways to explain by adding additional explanations to the XAI, and responsiveness through progressive disclosure by allowing user-customised information visibility. The validation of the XUI involved semi-structured interviews with eight professionals with a background in cardiology, analysed through thematic analysis to synthesize the findings.

Analysis

Key findings suggest significant insights regarding the XAI methods and the application of design principles to enhance interpretability. While issues were noted with the XAI methods themselves, employing design principles such as complementary naturalness, flexibility through multiple ways to explain, and responsiveness through progressive disclosure effectively enhanced interpretability. Interestingly, the principle of sensitivity to context and mind, although not initially implemented, was highlighted as fundamentally important through the analysis, leading to its proposed placement at the pinnacle of a restructured pyramid model of design principles.

The findings also underscored the essential role of involving stakeholders early in the development of XAI and XUI to ensure the design cycle is informed by end-user needs. Furthermore, feedback on the XUI's design elements indicated their utility primarily during initial interactions, with preferences for more streamlined and direct explanations in frequent use scenarios. Additionally, the research illuminated

the challenge of balancing the accurate explanation of AI processes with user expectations in medical contexts, suggesting a need for XAI designs that maintain transparency while adapting explanations to enhance decision-making efficacy without diluting the AI's reasoning integrity.

Conceptualising these insights, a framework for a design approach to XUI was proposed, consisting of four sequential phases: pre-XAI design, XAI design, pre-XUI design, and XUI design, recognizing XAI design as an integral phase to XUI development. The integration of the two propositions—the pyramid model and the design approach framework—is designed not only to align the XAI and XUI with users' needs but also to ensure that interactive design principles are effectively implemented. This approach aims to enhance the overall interpretability and functionality of the system, making it more user-friendly and efficient in meeting specific user requirements.

Future research

Practical recommendations highlight the need for future research, suggesting that future iterations could benefit from developing targeted strategies that address model understanding and decision-making efficiency separately, possibly through adaptive interfaces tailored to user interaction levels. Further in-depth exploration of (X)AI within the medical field is recommended, particularly through collaborations with medical experts to enrich the research and ensure that (X)AI applications are effectively integrated into medical practices. Additionally, engaging with domain experts from the start could help align XAI development more closely with end-user needs and practical domain-specific requirements, especially in deciding whether to prioritize AI's reasoning transparency or alignment with user expectations. The proposed design approach framework should serve as a guideline to facilitate this active engagement with relevant stakeholders. Lastly, the initial design of the XUI should be seen as a foundational step in an iterative process, with future projects encouraged to utilize feedback from initial evaluations to guide the development of subsequent iterations. This approach would allow for continuous refinement and adaptation of the XUI, enhancing its utility and relevance across various decision-support contexts.

Contents

Preface	i
Summary	ii
1 Introduction	1
1.1 Research background	1
1.1.1 AI developments	1
1.1.2 XAI	1
1.2 Research gap	2
1.3 Research questions	3
1.4 Research scope	3
1.5 Research Relevance	4
1.5.1 Scientific relevance	4
1.5.2 Societal relevance	4
1.6 Reading Guide	5
2 Research Approach	6
2.1 Research Methodology	6
2.2 Research Phases	7
2.2.1 Research Phase I: Identify	7
2.2.2 Research Phase II: Define	7
2.2.3 Research Phase III: Design & develop	8
2.2.4 Research Phase IV: Demonstration & Evaluation	8
2.2.5 Research Phase V: Communication	8
2.3 Research Flow Diagram	9
3 Literature	10
3.1 Interpretability	10
3.1.1 Human interaction	10
3.1.2 Bias and prior knowledge	11
3.1.3 Abductive reasoning	11
3.1.4 Causality	11
3.1.5 Constrastive Explanation	12
3.1.6 Inherent and extrinsic features	12
3.1.7 Explanation Selection	13
3.1.8 Explanation Evaluation	13
3.1.9 Multiple ways of explanation	14
3.1.10 Context dependent	14
3.1.11 Conclusion	18
3.2 Interaction	19
3.2.1 Explainable User Interfaces types	19
3.2.2 Explanatory goals in XAI	20
3.2.3 Interaction defined	20
3.2.4 Design Principles	22
3.2.5 Conclusion	23
4 Design	24
4.1 Use Case: FOKUS	24
4.1.1 ECG data	25
4.1.2 FOKUS XAI	26
4.1.3 Concerns and limitations	29

4.1.4	Selection of use case	29
4.2	Aligning literature with use case	30
4.2.1	Conceptual framework	30
4.2.2	AI in the medical domain	31
4.2.3	Theoretical Framework	32
4.2.4	Conclusion	36
4.3	XUI Design	37
4.3.1	XUI elements	37
4.3.2	Design principles applied	43
4.3.3	Conclusion	43
5	Validation XUI	44
5.1	Data Collection and Analysis method	44
5.2	Sampling Process	44
5.3	Interview Process	45
5.4	Interview data analysis	46
5.5	Results	48
5.5.1	Themes	48
5.5.2	Problems with (X)AI	48
5.5.3	Complementary naturalness	51
5.5.4	Responsiveness through progressive disclosure	52
5.5.5	Flexibility through multiple ways to explain	52
5.5.6	Environmental setting of the XUI	53
5.5.7	Diagnosis	54
5.5.8	Sensitivity to context and mind	54
5.5.9	XUI design	55
6	Research Findings & Discussion	56
6.1	XUI design	56
6.1.1	Design principles	56
6.1.2	Explanations elements	58
6.1.3	XAIs	59
6.1.4	Deployment of XUI	60
6.2	Proposition to Design Principles	60
6.3	Conceptual framework reflection	62
6.4	XAI design	64
7	Conclusion	65
7.1	Layered Knowledge Gap	65
7.2	Main research question	65
7.3	Limitations and future research	66
7.4	Reflections	67
	References	70
A	Interview Protocol	73
B	Code Analysis	75
B.1	Code book	75
B.2	Theme description and grouping	76
C	Overview XAI outputs	78
D	XUI pages	80

List of Figures

1.1	XAI and XUI concept (Jin et al., 2021)	2
2.1	Design Science Research Method (DSRM) visualisation (Peppers et al., 2007)	6
2.2	Multidisciplinary approach to literature review	7
2.3	Research Flow Diagram	9
4.1	Components of the ECG complex (Hampton & Hampton, 2019)	25
4.2	Standard 12 lead ECG display (Hampton & Hampton, 2019)	25
4.3	SHAP with positive and negative indications	27
4.4	LIME with varying colour scale	27
4.5	Grad-CAM outputs with different colour gradients	28
4.6	Conceptual framework	30
4.7	Theoretical framework diagram from literature results	33
4.8	Chosen XAI methods	37
4.9	Explanation element: Patient information	38
4.10	Explanation element: XAI method	39
4.11	Explanation element: Prediction score	39
4.12	Explanation element: XAI output	40
4.13	Explanation element: Prominent leads	40
4.14	XUI displaying explanation 1 for patient 1	41
4.15	Disclosure of element prediction score in open and closed state	42
4.16	Initial presentation of XUI displaying explanation 1 for patient 1	42
5.1	Qualitative Data Analysis Process	44
6.1	Proposed restructuring of the Design Principles by Chromik and Butz (2021)	61
6.2	Issues identified in the conceptual framework	62
6.3	Framework: Design approach for interactive XUI	63
C.1	Side by side view of the XAI outputs for two patients	79
D.1	XUI pages for patient 1	81
D.2	XUI pages for patient 2	83

List of Tables

3.1	A summary of the different factors of interpretability from the literature review	15
4.1	Interactivity Types and Linked Design Principles	34
4.2	Interpretability Factors and Their Corresponding Interactivity Types	35
5.1	Sampling Process	44
5.2	Interview Participants Overview	45
5.3	Process stages of a Thematic Analysis (Inspired by Braun and Clarke (2006))	46
5.4	Thematic Analysis Framework	47
5.5	Code frequency within themes	48
6.1	Feedback on the implementation of the design principles	56
B.1	Themes identified through theme analysis, the corresponding codes and description . . .	76

1

Introduction

1.1. Research background

1.1.1. AI developments

In recent years, machine learning (ML) has seen a tremendous amount of development and growth, resulting in a new wave of Artificial Intelligence (AI) algorithms. These algorithms offer significant benefits to science and industry with their enhanced predictive capabilities and accuracy. However, this growth has also resulted in a deeper level of complexity that affects both the designers and users of these algorithms.

One of the major challenges posed by ML algorithms is their inability to explain their autonomous decisions to human users and why they are considered "black boxes". Due to their complex nature, users are unable to understand the reasoning behind a model, adding to the complexity of these algorithms. This challenge is even greater in the case of deep learning models, which are attributed as a large part of the recent AI developments. These models often outperform traditional ML algorithms, but are even less explainable, making them even more complex and challenging to understand (Ahmad et al., 2019).

The lack of explainability in these new ML models can be a significant concern, particularly in fields which involve human welfare such as healthcare, finance and law, where the consequences of incorrect decisions can be severe. In these industries, it is crucial to have a clear understanding of how a model arrived at its decision in order to ensure that the decision can be deemed valid and trustworthy. However, due to the complex nature of ML models and their inability to explain their reasoning, they lack transparency which affects their perceived trustworthiness by the users. This is a major concern as it can be a hindrance to their adoption in these and many other fields (Rajpurkar et al., 2022).

1.1.2. XAI

To address concerns surrounding the opacity of artificial intelligence systems, the field of Explainable Artificial Intelligence (XAI) has emerged as a crucial area of study. This discipline focuses on enhancing the interpretability and accessibility of machine learning models for a broader audience. The primary goal of XAI is to promote transparency and foster trust by clearly elucidating the decision-making processes of AI systems (Barredo Arrieta et al., 2020). This is achieved through various methods, such as intuitive visualisations that depict the model's reasoning, natural language explanations that simplify complex mechanisms, and other techniques designed to make AI's operations transparent (Gunning et al., 2019).

Furthermore, XAI is instrumental in identifying and addressing biases and errors within AI models, thereby enhancing their fairness and accountability. As AI applications become increasingly integrated across diverse sectors, understanding these systems is essential not only for developers but also for end-users who rely on AI for critical decision-making. XAI serves to ensure that AI decisions are based on transparent and justifiable processes, making these systems more reliable and trustworthy. Figure

1.1 illustrates how XAI achieves this, in contrast to conventional AI approaches, by highlighting the move towards more transparent AI practices where both users and practitioners can understand and trust the underlying logic of AI outputs.

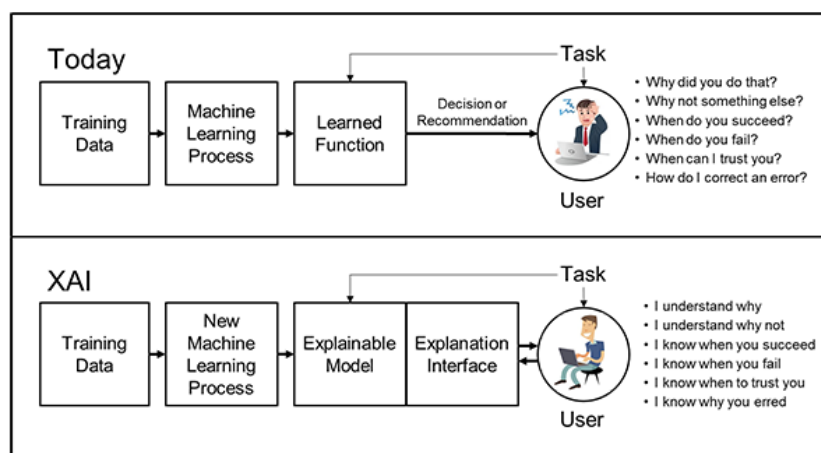


Figure 1.1: XAI and XUI concept (Jin et al., 2021)

1.2. Research gap

In the field of XAI, two essential concepts, interpretability and explainability, focus on making the inner workings of AI systems understandable to humans. Interpretability involves breaking down complex AI decisions into simpler, human-understandable terms. Despite extensive discussions in the field, a universally accepted definition of interpretability is lacking, indicating that individuals' understanding of this concept varies based on their background and the explanation methods utilised (Adadi & Berrada, 2018; Ribeiro et al., 2016). Nonetheless, the importance of interpretability is recognised for its role in deepening our understanding of AI systems and enhancing trust in these technologies (Barredo Arrieta et al., 2020).

Research in XAI has largely focused on those with a technical background, often neglecting the broader spectrum of end-users who lack such expertise. This oversight underscores a critical challenge: developing explanations that accommodate a diverse audience with varying levels of understanding. Addressing this challenge is critical, particularly as AI technologies gain traction in sectors critical to societal welfare (Bhatt et al., 2019; Jin et al., 2021).

User interfaces are identified as crucial tools for enhancing the accessibility of AI explanations. By designing intuitive and user-friendly interfaces, there can be significant improvement in individuals' interpretability of AI decision-making processes. This improvement in interpretability is vital for building user trust and encouraging the adoption of AI across different fields (Chromik & Butz, 2021; Doshi-Velez & Kim, 2017; Jin et al., 2021).

Furthermore, interactivity can also play a pivotal role in enhancing interpretability in XAI. Interactive user interfaces allow users to engage directly with the AI's decision-making process, offering a hands-on approach to understanding complex information. By interacting with the AI system, users can explore different scenarios, ask questions, and see the impact of changes. This interaction not only makes the AI's operations more transparent but also allows users to learn at their own pace, catering to various learning styles and levels of expertise. Incorporating interactivity into XUIs could bridge the gap between complex AI functionalities and user comprehension, making interpretability a more dynamic and personalised experience (Abdul et al., 2018; Chromik & Butz, 2021; Madumal et al., 2018).

While there have been efforts to align XAI with user needs through frameworks (Adadi & Berrada, 2018; Gunning et al., 2019; Jin et al., 2021), a notable absence exists in the area of design principles tailored for XUIs. This shortfall is particularly evident when it comes to the practical deployment of these principles, especially for meeting the needs of diverse user groups. Current frameworks tend

to oversimplify user needs, offering broad generalizations without providing detailed approaches for real-world application (Chromik & Butz, 2021). The gap can be attributed to several factors. One of the primary factors could be the time constraints associated with research papers, which may not allow for an extensive exploration of proposed findings and guidelines. This constraint is frequently acknowledged in the conclusions and future research sections of such papers (Chromik & Butz, 2021; Jin et al., 2021). Another contributing factor to the gap is the continuously evolving nature of XAI, as it is a relatively new field with areas still unexplored. Additionally, XAI is a multidisciplinary domain that necessitates expertise from various fields, including computer science, statistics, social science, and human factors. The integration of these diverse perspectives can be challenging.

This underlines the need for dedicated research aimed at developing practical, interpretable XAI interfaces with the help of refined design principles. The research should prioritise understanding of the distinct preferences and needs of end-users. The objective is to improve design principle enhancing design principles to create interpretable interfaces that are interactive, user-centric and easy to navigate for individuals from all backgrounds and levels of expertise.

1.3. Research questions

Building on the insights gathered about critical issues within the field of XAI interpretability, two primary knowledge gaps have been identified that necessitate further exploration. The first gap highlights the current explanation methods which, although rich in various strategies to enhance interpretability, primarily cater to users with technical expertise, neglecting a wider audience that lacks such expertise.

The second gap pertains to the practical application and development of user-centric design principles for XUIs. The framework developed by Chromik and Butz (2021) introduces innovative interactive design principles for enhancing XAI systems, yet the practical application of these principles in real-world settings where actual users interact with AI remains significantly under-explored. This deficiency points to a need for rigorous empirical research that focuses on deploying and testing these principles in real-world settings to assess their effectiveness and adaptability.

Given these identified gaps, the main research question has been formulated as follows:

How can interactive design principles be strategically applied to Explainable User Interfaces (XUI) to enhance the interpretability of Explainable Artificial Intelligence (XAI) for end-users?

In order to answer the main research question, the following sub-research questions have been formulated as follows :

1. What are the factors that make an explanation interpretable for end-users?
2. How does interaction influence the interpretability of end-users?
3. How can interpretability and interactivity be linked to design principles for an XUI?
4. How can interactive design principles be applied to design an XUI for end-users?

1.4. Research scope

To maintain a clear focus within the designated subject matter, it is essential to define several key dimensions within the scope of the research. This includes specifying the deliverable, outlining the subject matter, and establishing the boundaries of the study.

- The primary aim of this research is to develop and evaluate the XUI and its various design principles, rather than delving into the broader UI design. This study is specifically targeted at how XUI can enhance the interaction between users and AI systems by making the outputs of AI more transparent and comprehensible. The focus is on assessing and refining design principles that facilitate effective human-AI collaboration, improving the overall user experience and usability of AI systems.

- This research will not focus extensively on the development of the underlying XAI systems themselves. XAI involves creating AI systems that can explain their decision-making processes in a transparent and interpretable manner. Designing XAI requires specialised knowledge and often involves complex technological development, which falls outside the scope of this study. Instead, the emphasis will be on how the XUI interfaces with existing XAI systems to improve user interactions.
- The deliverable of this research is a prototype XUI that will be based on a specific use case provided by Fraunhofer FOKUS, an institute for open communication systems. FOKUS is working on developing wearable emergency medical devices that monitor an individual's vital signs and utilise AI to detect abnormalities. The XAI developed by FOKUS for this application will serve as the basis for the XUI in this study and will be discussed during the design phase.
- While the objective of this research is to enhance interpretability, the thesis will not delve into methods for measuring interpretability, as this topic warrants an entire thesis of its own. Instead, a more pragmatic approach is adopted to assess the effectiveness of the XUI. This involves conducting semi-structured interviews to determine whether the XUI has improved users' understanding of XAI system outputs, thereby enhancing their overall interpretability.

1.5. Research Relevance

1.5.1. Scientific relevance

The current state of research in explainable artificial intelligence (XAI) exhibits a strong theoretical foundation, yet there is a noticeable disconnect when it comes to the practical application of these theories, especially in non-technical user contexts (Chromik & Butz, 2021). This research is pivotal as it seeks to bridge this gap by applying interactive design principles within explainable user interfaces (XUIs) directly to real-world scenarios. By focusing on empirical testing and refinement of these principles, the study aims to transform theoretical concepts into effective, user-friendly tools that enhance the interpretability of AI systems across diverse societal groups.

This effort is crucial for advancing the practical deployment of XAI, ensuring that theoretical advancements lead to tangible improvements in how AI systems are perceived and interacted with by everyday users (Doshi-Velez & Kim, 2017; Gunning & Aha, 2019; Jin et al., 2019). The research not only enriches the academic landscape by providing validated models for user interaction but also sets a precedent for future studies to prioritise the practical implications of their findings. Such a focus ensures that advancements in AI and machine learning are more readily understandable and accessible, thereby increasing the overall acceptance and ethical integration of these technologies into daily life.

1.5.2. Societal relevance

In societal terms, the relevance of this research is underscored by recent regulatory initiatives, notably the European Union's Artificial Intelligence Act proposed in April 2021 (Commission, 2021). This groundbreaking regulation adopts a risk-based approach, imposing strict requirements for transparency, accountability, and human oversight, especially in high-risk AI applications. The act stipulates that XAI techniques must be employed in specific applications to facilitate the explanation of AI system decisions to users. This underscores the growing necessity for transparency and accountability in AI systems, highlighting the crucial role of enhanced interpretability.

By improving how AI systems are explained and understood, this study supports the ethical deployment of AI, addressing critical concerns such as bias and fairness while promoting inclusivity (Gunning & Aha, 2019). Making complex AI decisions understandable and relatable to a broad audience is crucial, especially as AI technologies increasingly permeate sectors with substantial societal impacts, such as healthcare, finance, and public services (Bhatt et al., 2019; Jin et al., 2021). Enhancing the interpretability of AI not only fosters broader public engagement and trust but also facilitates a more informed dialogue about the ethical and practical implications of AI in daily life. This contributes to a safer, more equitable integration of AI technologies into society, ensuring that advancements in AI align with public interests and regulatory standards, thereby promoting positive societal welfare (Doshi-Velez & Kim, 2017; Rajpurkar et al., 2022).

1.6. Reading Guide

This thesis begins by defining the scope of the problem and identifying the research gap that subsequent chapters will address. Chapter 2 describes the methodological framework employed in this study, tracing the journey from the initial identification of design principles to their final evaluation. Chapter 3 offers an in-depth literature review focusing on interpretability, types of interactivity, and design principles within the realm of XAI. Chapter 4 introduces the FOKUS use case, aligns it with the findings from the literature review, and constructs both a conceptual and a theoretical framework. This chapter also details the design of the XUI by applying selected interactive design principles. Chapter 5 validates the XUI through thematic analysis of interview data, demonstrating the implementation of design principles and their impact on end-user interpretability. Chapter 6 evaluates the effectiveness of these principles, summarising the findings and discussing their implications for future XUI design. It also proposes potential improvements based on the insights gathered throughout the research. The thesis concludes with Chapter 7, which synthesises the findings from all chapters to provide a comprehensive answer to the primary research question. This final chapter integrates responses to all sub-questions, discusses the limitations of the study, and suggests directions for future research. This structured approach ensures a thorough exploration of how interactive design principles can be effectively applied to develop an XUI that meets the diverse needs of its users

2

Research Approach

2.1. Research Methodology

This research will be carried out by applying the Design Science Research Method (DSRM) (Peppers et al., 2007). The research methodology emphasises the development of innovative artifacts or solutions to practical problems, while also advancing knowledge in a particular domain. DSRM is used in various fields, including information systems, engineering, and healthcare. DSRM was introduced as a way of addressing the challenges of designing and implementing information systems in complex and dynamic environments. DSRM involves an iterative process of problem identification, design, implementation, and evaluation, with the ultimate goal of producing a novel artifact that can be tested and refined in real-world settings. The novel artifact in this case would be the deliverable of the XUI targeted towards end-users to increase their interpretability in XAI. The method and its activities can be seen in figure 2.1

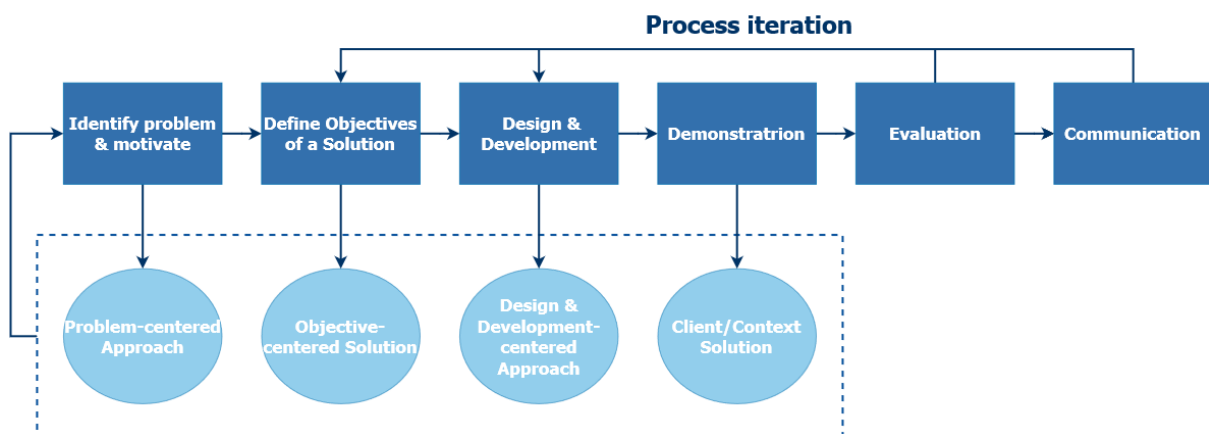


Figure 2.1: Design Science Research Method (DSRM) visualisation (Peppers et al., 2007)

The Design Science Research Methodology (DSRM) involves six activities depicted in figure 2.1. Initially, the "Identify Problem & Motivate" activity pinpoints a practical problem or opportunity that necessitates a solution, ensuring it is relevant to the field of study and aligned with overarching research goals. Following this, the "Define Objectives of a Solution" activity aims to design an artifact that addresses the identified problem, informed by outcomes from multiple literature reviews that serve as a basis

for designing the XUI. The "Design & Development" activity then involves the actual implementation of the XUI prototype. Subsequently, the "Demonstration & Evaluation" activity showcases the artifact to the relevant target group through semi-structured interviews, evaluating the results to assess the effectiveness of the solution in addressing the initial problem. Finally, the "Communication" activity disseminates the evaluation results to stakeholders, ensuring the artifact meets their needs, culminating in the finalisation and conclusion of the research findings in a completed master thesis.

2.2. Research Phases

2.2.1. Research Phase I: Identify

In the initial "Identify" phase of the research, the study began by pinpointing a real-world problem or opportunity requiring a solution, setting the foundational direction of the investigation. This phase was crucial in ensuring that the identified problem was not only relevant to the field of study but also aligned with the overarching goals of the research. It highlighted a significant gap in the field of XAI: the pressing need for focused research aimed at enhancing interpretability through the use of interactive user interfaces. This discovery set the stage for subsequent phases, directing the research towards addressing this key challenge and contributing valuable insights to the domain of XAI. The main research question formulated based on this research gap is :

How can interactive design principles be strategically applied to Explainable User Interfaces (XUI) to enhance the interpretability of Explainable Artificial Intelligence (XAI) for end-users?

2.2.2. Research Phase II: Define

The following sub-research questions will be answered during this phase :

What are the factors that make an explanation interpretable for end-users?

How does interaction influence the interpretability of end-users?

How can interpretability and interactivity be linked to design principles for an XUI?

To achieve the aims of these sub-research questions, insights from the social sciences, Human-Computer Interaction (HCI), and other relevant fields will be gathered and analysed. This multidisciplinary approach is crucial for exploring and synthesising literature on theories, frameworks, and studies pertinent to the interpretability of explanations for end-users. Understanding what makes an explanation interpretable and how interaction affects this interpretability is essential for enhancing the effectiveness of Explainable Artificial Intelligence (XAI) systems. This approach is visualised in Figure 2.2.

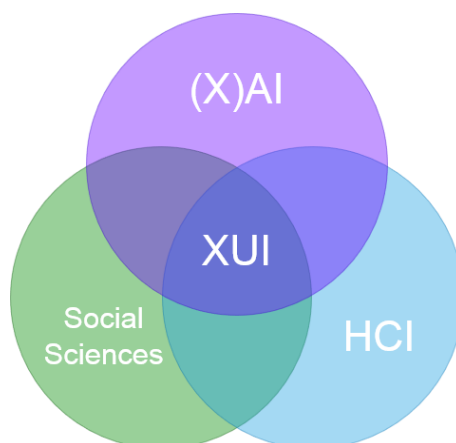


Figure 2.2: Multidisciplinary approach to literature review

For the first question, the focus will be on identifying factors that contribute to making explanations clear and understandable to end-users, examining how various elements of an explanation impact its interpretability and usefulness in the context of XAI systems.

The second question explores the role of interaction in the interpretability process, seeking to understand how engaging with the system and its explanations influences end-users' ability to comprehend and utilise the information provided by XAI systems.

The third question aims to synthesise the findings from the first two questions, linking and translating these insights into actionable design principles. This synthesis will establish a theoretical foundation for these principles based on the reviewed literature.

Given the broad exploration of interpretability and user interaction within different fields, a comprehensive literature review will serve as the most suitable approach to address these sub-research questions. This review will be conducted as part of the "Define Objectives of a Solution" phase in the Design Science Research Methodology (DSRM), aiming to extract insights that can guide the enhancement of interpretability in XAI systems.

2.2.3. Research Phase III: Design & develop

In this phase, factors identified from the previous sub-research questions will be assessed and applied to the specified use case and XUI. A theoretical and conceptual framework will be derived from these findings and tailored to fit the use case. Subsequently, the XUI will be designed based on the literature review, expert advice, and creative input. Additionally, an interview protocol will be developed to be used after implementing the XUI, enabling the measurement of the applied design principles for effectiveness in interpretability.

The phase focuses on the development of the XUI, beginning with the creation of rough prototypes to integrate the theoretical insights and creative ideas into a mind-map. This will progress to the design of mock-ups in Figma, and eventually lead to the development of a working prototype. This research phase aims to answer the sub-research question:

How can interactive design principles be applied to design an XUI for end-users?

2.2.4. Research Phase IV: Demonstration & Evaluation

This sub-research question focuses on assessing the impact of the interactive XUI on end-users' ability to interpret XAI outputs, aiming to understand how the XUI influences their interpretability. This inquiry builds on insights from the initial sub-research question that identified key factors affecting interpretability.

The exploration of this question is scheduled for the "Demonstration & Evaluation" phase, following the design, implementation, and case study demonstration of the XUI. This approach allows for a practical assessment of the XUI's effectiveness in real-world applications, aiming to close the gap in current understanding.

The objective is to gather detailed evidence through semi-structured expert interviews, where experts are introduced to the XUI. Analyses of these interview data will subsequently be conducted to comprehensively address the main research question. The findings from this segment of the study will significantly enrich the overall analysis and conclusions, elucidating the role of interactive XUIs in enhancing the interpretability of XAI for end-users.

2.2.5. Research Phase V: Communication

The final phase, "Communication," is dedicated to discussing the outcomes of the research and addressing the main research question. In this stage, the findings are thoroughly analysed to draw meaningful conclusions about the effectiveness of the interactive design principles on the interpretability of XUIs, focusing on how well they meet end-users' needs. Additionally, it involves an honest examination of the study's limitations, acknowledging areas where the research might have fallen short or aspects that were beyond the scope. This sets the stage for future research, indicating paths that subsequent studies could follow to delve deeper into unresolved questions or tackle new challenges that have emerged. Furthermore, this stage includes reflection on the entire research process, evaluating what was learned and how these lessons can guide future projects.

2.3. Research Flow Diagram

To illustrate the research approach adopted in this thesis, a research flow diagram has been developed, outlining each phase of the Design Science Research Method (DSRM). This diagram details the corresponding activities of each phase, including the research methods employed, specific research activities undertaken, the chapters these activities are documented in, and the research questions they address. This comprehensive visual representation aids in understanding the systematic progression of the research and is displayed in Figure 2.3.

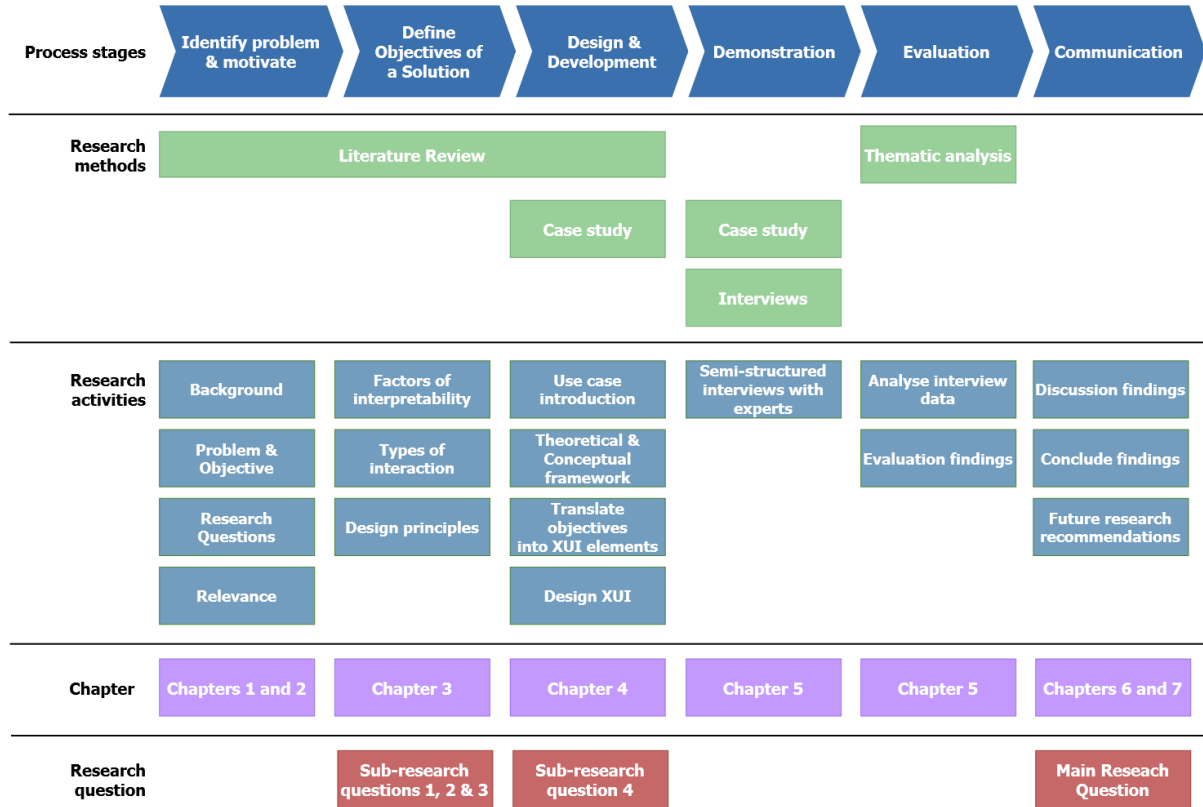


Figure 2.3: Research Flow Diagram

3

Literature

This chapter is dedicated to addressing sub-research questions 1 and 2, focusing on identifying factors associated with interpretability in humans and exploring the role of interactivity in enhancing this interpretability. Additionally, it delves into the design principles specific to the domain of XUI design. The research will be conducted through an extensive literature review, drawing insights from existing literature within these areas. The chapter concludes by summarising the findings from the literature, effectively answering the two sub-research questions.

3.1. Interpretability

The literature in this field is extensive, diverse, and often complex. In this section, the factors related to interpretability that are most frequently discussed in the literature have been selected and elaborated upon. Interpretability in the context of (X)AI refers to the extent to which the internal mechanics of a machine or deep learning model can be understood by humans. It is a crucial aspect of (X)AI that allows users to comprehend and trust the decisions made by these systems. Measuring interpretability involves assessing how well a person can predict the model's response to changes in input or understand the reasoning behind its decisions, the latter being the case for this research (Barredo Arrieta et al., 2020; Gunning et al., 2019; Miller, 2019).

3.1.1. Human interaction

De Graaf and Malle (2017) argue that people tend to attribute human-like traits to artificial agents and therefore expect these systems to provide explanations in a human-like manner. This underscores the need for autonomous systems to deliver explanations that resonate with natural human ways of explaining and interpreting actions. Similarly, Jin et al. (2021) emphasise that for AI systems to be embraced by non-technical end-users, these systems must align their explanations with intuitive human reasoning and decision-making processes. This alignment is crucial for fostering user understanding and acceptance.

Further expanding on this concept, Holzinger and Muller (2021) point out that humans have an inherent desire for continuous feedback, which in face-to-face interactions, is often conveyed through facial expressions. The incorporation of "emotional interfaces" in AI systems could potentially mimic this aspect of human interaction, enhancing the user experience by making interactions seem more natural and responsive.

Additionally, Miller (2019) conceptualises explanations as a form of social interaction that involves two roles: the explainer and the explainee. In the context of AI, the artificial agent acts as the explainer, while the human user is the explainee. This dynamic underscores the importance of understanding that explanations are not just about transferring knowledge but also about fitting into the social and cognitive contexts of the explainee, taking into account both the explainer's and the explainee's beliefs and expectations. This perspective highlights the relational aspect of explanations and suggests that effective communication in AI systems should consider the social dynamics between human and

machine.

3.1.2. Bias and prior knowledge

People inherently carry biases and social expectations that influence how they receive and interpret explanations. According to Miller (2019), these biases and expectations can actually enhance interactions with explainable artificial intelligence (XAI) by making the explanations more relatable or understandable to users. However, biases are a double-edged sword; while they can facilitate understanding when explanations align with pre-existing beliefs, they can also obstruct it when they do not.

Bias impacts explanation by shaping an individual's ability to link causes with effects, which is often rooted in their prior beliefs. When explanations are founded on accurate or truthful beliefs, they can significantly aid in understanding complex causations. Conversely, explanations based on false beliefs or inaccurate information can reinforce misunderstandings and spread misinformation, as noted by Lombrozo (2007).

Furthermore, the bias of the explainer themselves can influence the explanations they provide. Lombrozo (2007) points out that establishing clear boundaries and understanding the inherent biases in explanations can help manage and possibly mitigate their impact on the reasoning processes of the explainee. This approach can guide the design of more effective and impartial XAI systems, ensuring that explanations serve to clarify rather than confuse.

3.1.3. Abductive reasoning

Explanation is intimately linked with abductive reasoning, a cognitive process distinct from but sometimes confused with induction. Charles Peirce, who first distinguished abduction as a unique form of reasoning, described it as moving from effect to cause, unlike deduction. Peirce highlighted the contrast between inductive reasoning, which validates a hypothesis through scientific experimentation, and abductive reasoning, which formulates a hypothesis to explain an observed phenomenon.

Abduction involves crafting a plausible hypothesis or explanation for a given observation, even in the absence of full evidence. It is particularly useful in situations where multiple potential explanations exist for a single observed phenomenon. This type of reasoning is often seen as making a "best guess" or a plausible inference based on the information at hand, guiding further investigations that may confirm or refute the hypothesis.

This process, termed "inference to the best explanation" by Harman (1965), involves advocating for a hypothesis or theory as the most suitable fit for the observed data. It is supported by demonstrating that the selected hypothesis or theory provides a more comprehensive and satisfactory explanation compared to other alternatives. In this way, abductive reasoning not only aids in understanding complex phenomena but also drives the scientific inquiry forward by suggesting new areas for exploration and confirmation.

3.1.4. Causality

In both philosophy and psychology, there is a shared understanding that causality plays a pivotal role in explanations. Typically, providing an explanation involves identifying and attributing causes to comprehend why or how certain events, phenomena, or behaviors occur. Causality is fundamentally linked to the process of making sense of the world, as it helps elucidate the underlying reasons behind observable outcomes.

For instance, consider the scenario where a shadow is cast by a flagpole in the presence of the sun. In this example, the flagpole and the sun are deemed the causal agents explaining the presence of the shadow, whereas the shadow itself cannot explain the existence of the flagpole or the sun. This differentiation highlights the nature of causation over mere correlation, emphasizing a "would have" relationship that signifies direct causality (Keil, 2006).

The concept of causality also incorporates the notion of counterfactual thinking, as outlined by Lewis (2013). Counterfactual reasoning involves hypothesizing about alternative outcomes by questioning what might have occurred if a specific event (Event A) had not taken place, and considering the potential impact on another event (Event B), assuming all other conditions stayed constant. This type of thinking

is often employed at a psychological level to clarify and emphasise causal relationships, despite its complex philosophical undertones (Keil, 2006).

However, merely presenting a causal chain does not automatically provide a complete explanation. Such causal attributions serve as tools that individuals might use to derive their own explanations. It is argued, particularly in the context of AI models, that expecting a layperson to interpret a complex causal chain is unrealistic, regardless of the presentation format (Miller, 2019).

3.1.5. Contrastive Explanation

Research suggests that when individuals provide explanations for events, they often frame these in terms of causality relative to an alternate scenario that did not occur. This method, often encapsulated in queries such as "Why P rather than Q?", where P is the event being explained and Q is a counterfactual contrast case, highlights differences between two possibilities—one that occurred and one that did not, even if Q is not explicitly mentioned. This approach is known as a contrastive explanation, which is detailed by Miller (2019) as focusing on the differences between similar events that result in different outcomes. For instance, in comparing two medical treatments with differing outcomes, a contrastive explanation would explore factors like dosage or side effects to determine what influenced these different outcomes.

Contrastive explanations differ significantly from counterfactual explanations in causality. While counterfactuals engage with hypothetical scenarios to explore what might have happened had a specific event not occurred, contrastive explanations deal directly with real events, comparing them to discern the causal differences.

Lipton (1990) further refines this idea by stating that contrastive explanations particularly emphasise the distinctions between an actual event (P) and a hypothetical alternative (Q). He notes that contrastive explanations are generally easier to generate than complete explanations for factual inquiries about P because they focus solely on the differences rather than the full causality. For example, when asked "Why is image J labelled as a Beetle rather than a Spider?", a contrastive explanation could point out the presence of six legs—a characteristic of beetles—without needing to discuss other features like eyes or wings. In contrast, a non-contrastive query such as "Why is image J a spider?" would require a comprehensive explanation that includes every trait that categorises an arthropod as a spider.

Lipton argues that contrastive explanations simplify the explanatory process by focusing only on critical differences between two cases, making them particularly effective for both human understanding and computational applications where full causal analysis may be impractical or unnecessary. This targeted approach not only streamlines the explanation process but also pinpoints the specific areas of a model or scenario that a questioner may find confusing. By directly addressing these areas, contrastive explanations enhance clarity and effectiveness, making complex information more accessible and understandable. This method proves invaluable in both academic and practical settings, facilitating clearer communication and more effective understanding of complex phenomena (Miller, 2019).

3.1.6. Inherent and extrinsic features

Research by Prasada and Dillingham (2006) and Prasada (2017) explores how abductive reasoning tends to prioritise certain factors over others, highlighting the cognitive processes involved in categorizing and explaining properties of objects. Prasada posits that recognizing something as an instance of a specific kind and explaining its characteristics based on its classification are not separate activities but rather a unified cognitive process. This approach helps to understand how people differentiate between types of properties within a category: k-properties and t-properties.

K-properties, or principled connections, are inherent properties that are essentially tied to the nature of the kind itself. T-properties, or factual connections, on the other hand, are extrinsic and not inherently connected to the essence of the kind. Prasada's experiments demonstrate that in formal modes of explanation, properties that relate to the kind or category of an object (k-properties) are considered more relevant and provide stronger explanations than those that pertain to extrinsic properties (t-properties). Moreover, explanations that utilise an object's category to explain why it possesses certain properties are found to be more effective for k-properties than for t-properties.

For instance, in explaining why a bird can fly—a k-property—citing that it belongs to the category of

birds is more effective than referring to its color or size, which are t-properties. This distinction is crucial in abductive reasoning, especially in classification tasks where understanding the nature of k-properties can significantly enhance the accuracy and relevancy of the categorization and subsequent explanations tasks (Miller, 2019).

3.1.7. Explanation Selection

When providing explanations for events, people typically do not enumerate all the causes involved. Instead, they select and emphasise what they perceive as the most relevant causes or factors. This selection process is influenced by various factors that dictate how humans prioritise different aspects of an explanation:

- **Contrastive explanations** People often frame explanations in a contrastive manner, distinguishing between a specific event and a potential alternative that did not occur. This method highlights the most critical aspects of the actual event by contrasting it with a counterfactual scenario (Lipton, 1990).
- **Abnormal events** According to D. J. Hilton and Slugoski (1986), explainers leverage their perceived background knowledge along with that of the explainees to highlight conditions considered abnormal. This focus on abnormality helps to clarify why certain events occur outside the expected norm.
- **Intention and function** Explanations that consider the intention behind an action or the function of an object are often seen as more satisfactory than those that merely describe abnormalities. This preference indicates that explanations which attribute a purpose or function to an event or object resonate more strongly, providing a higher level of explanatory power (D. Hilton et al., 2005; Lombrozo, 2010).
- **Necessity, sufficiency and robustness** These criteria help determine the strength of an explanation. An event that is necessary and sufficient for an outcome is often considered a robust explanation because it comprehensively accounts for the occurrence without requiring additional factors (Lipton, 1990; Lombrozo, 2010; Woodward, 2006).
- **Responsibility** The concept of responsibility plays a crucial role in causal selection, where an event deemed more responsible for an outcome is likely to be judged as a better explanation compared to others. This ties closely to necessity, as responsibility seeks to assign a 'degree of necessity' to causes. An event that is fully responsible for an outcome is viewed as a necessary cause. This aspect of explanation selection also intersects with the discussion of intrinsic versus extrinsic properties, where intrinsic properties (akin to necessary causes) are often deemed more explanatory than extrinsic ones (Miller, 2019).
- **Preconditions, failure and intentions** These elements also contribute to the selection of explanations. Preconditions and failures set the stage for an event, while intentions provide a deeper insight into the motives behind actions, enhancing the explanatory depth (Leddo et al., 1984; McClure & Hilton, 1997).

3.1.8. Explanation Evaluation

Various factors contribute to how humans evaluate explanations, each influencing the perceived quality and effectiveness of the information provided. These factors include:

- **Unsatisfying nature of probabilities** Explanations that rely solely on statistical probabilities are often found unsatisfying because they do not provide a causal connection. People generally seek causality in explanations rather than mere statistical relationships, which only show patterns or associations without implying causation. Josephson and Josephson (1996) discuss how statistical relationships alone may not satisfy the human need for understanding causality. Furthermore, McClure (2002) suggests that the quality of explanations is not judged primarily on their probability but rather on their ability to explain causal behavior in a pragmatic way. The interpretation of probabilities also poses challenges, as it is influenced by how information is framed; people tend to be risk-averse or risk-taking depending on whether the information is framed positively or negatively (Miller, 2019).

- **Coherence, simplicity, and generality** Thagard (1978) Theory of Explanatory Coherence highlights that for explanations to be accepted, they must be coherent with an individual's existing beliefs and simplistic in nature. The theory details seven foundational principles emphasising that explanations should seamlessly integrate with pre-existing knowledge and prioritise simplicity and generality. This means that explanations citing fewer causes and covering more events are preferred. Empirical support for Thagard's theory is evident in other studies (Read & Marcus-Newhall, 1993), where participants favored broader, simpler explanations, such as attributing multiple symptoms to a single cause like pregnancy over multiple, independent causes. This preference has significant implications across various fields including education, communication, and artificial intelligence, suggesting that effective explanations should align with the audience's knowledge and simplify complex information for better understanding and acceptance.

3.1.9. Multiple ways of explanation

The integration of multiple explanation methods is crucial for enhancing user understanding of machine learning (ML) models, as different individuals comprehend information in varied ways. Chromik and Butz (2021) advocate for the use of diverse explanatory approaches to cater to these differences, suggesting that understanding can be significantly enhanced when users are allowed to interact with and influence the inputs of a machine, thereby exploring alternative explanations and outcomes. Miller (2019) adds that people often seek explanations that explore the causal relationships in hypothetical or "what-if" scenarios, which can help clarify how different inputs might alter outcomes.

Expanding on the idea of using varied modalities for explanation, Madumal et al. (2018) discussed the effectiveness of multimedia narratives in complex communication scenarios. This approach tailors the mode of explanation—whether visual, textual, or auditory—to the context of the information and the background of the audience, thereby optimising engagement and trust. The level of detail in these explanations is adjusted based on the cultural and educational background of the audience, aiming to manage trust effectively and convey any uncertainties clearly.

Furthermore, the design of an intuitive and context-sensitive Explainable User Interface (XUI) is highlighted as essential for successful model interpretation. The XUI should present reasons behind AI decisions in formats that best suit the domain and the user's expertise—be it through images, text, or other mediums. Simple user interfaces may often be insufficient for dealing with complex system representations. The EUCA framework mentioned by Jin et al. (2021) supports this context-based approach, noting that the depth of interpretation required can vary significantly across different datasets and depend on the end-users' domain knowledge, emphasising the need for adaptable and user-centric explanation strategies.

3.1.10. Context dependent

The issue of tailoring explanations in machine learning (ML) applications is complex and varies significantly depending on the context, particularly when focusing on the diverse target group of "end-users." Within this group, users may possess varying degrees of expertise and experience related to both ML and the specific domain of application. For non-experts who lack familiarity with both the domain and ML, simple and comprehensible explanations are crucial. Conversely, domain experts, such as medical professionals who may be well-versed in their field but less so in ML, often require more detailed and nuanced explanations that respect their knowledge level. For example, such experts might find overly simplistic or vague explanations unsatisfactory or even offensive.

Miller (2019) emphasises that different users expect different types of explanations, underscoring the need to tailor communication to the audience's specific needs and questions. This customization extends to the method of explanation, which might involve varying the complexity of the information presented based on the user's familiarity with the subject matter. For instance, a heat map might be an effective way to visualise model gradients for image data but less so for tabular or textual data.

The essential building blocks for crafting effective explanations in ML can be summarised by addressing three critical questions:

- **What to explain?** Determining the type of data or feature that needs explanation is crucial. This involves understanding the aspects of the model or its outputs that are most relevant to the user's

needs and interests.

- **How to explain?** Choosing the appropriate method or communication tool is vital. Whether through textual descriptions, visual aids like graphs or heat maps, or interactive interfaces, the explanation method should align with the type of data and the user's ability to interpret that data.
- **Whom to explain to?** Identifying the target group is essential. Tailoring explanations to suit the knowledge level and expectations of the specific audience—whether they are ML novices, domain experts, or somewhere in between—ensures that the explanations are both useful and meaningful.

Addressing these questions effectively requires a nuanced understanding of both the audience and the data, ensuring that explanations enhance user understanding and trust in ML models.

Table 3.1: A summary of the different factors of interpretability from the literature review

Authors & year	Interpretability factor	Description
(De Graaf & Malle, 2017) (Jin et al., 2021) (Holzinger & Muller, 2021) (Miller, 2019)	Human Interaction	<ul style="list-style-type: none"> - People expect AI to provide explanations similar to how humans would, aligning with intuitive human reasoning for better acceptance. - Incorporating "emotional interfaces" in AI can mimic human expressions for more natural and engaging interactions. - Explanations within AI should consider the social and cognitive contexts of users, reflecting a dynamic between explainer (XAI) and explainee (human).
(Miller, 2019) (Lombrozo, 2007)	Bias And Prior Knowledge	<ul style="list-style-type: none"> - Biases and social expectations can enhance XAI interactions by making explanations relatable but can also obstruct understanding if they contradict pre-existing beliefs. - Bias shapes how individuals link causes and effects, aiding understanding when based on accurate beliefs, but potentially spreading misinformation when based on inaccuracies. - The biases of explainers can influence the explanations they provide, necessitating clear boundaries and understanding of these biases to ensure explanations clarify rather than confuse.
(Harman, 1965)	Abductive reasoning	A cognitive process where one infers the most likely cause or explanation for an observed phenomenon based on the best available evidence.

<p>(Keil, 2006)</p> <p>(Lewis, 2013)</p> <p>(Miller, 2019)</p>	<p>Causality</p>	<ul style="list-style-type: none"> - Causality is central in explanations, helping to identify causes that clarify why events occur, thus aiding in understanding the world around us. - Counterfactual reasoning involves hypothesising about what might have happened if conditions were different, helping to clarify causal relationships. - Presenting complex causal chains in AI models might be challenging for laypersons to interpret, indicating the need for simpler explanations .
<p>(Miller, 2019)</p> <p>(Lipton, 1990)</p>	<p>Contrastive Explanation</p>	<ul style="list-style-type: none"> - Examines why one event occurred instead of another, focusing on causality relative to an alternative scenario. - By highlighting only the differences between two scenarios, contrastive explanations simplify understanding, making them practical for both academic and computational applications. - Unlike counterfactuals that consider hypothetical alternatives, contrastive explanations deal with real events and discern causal differences.
<p>(Prasada & Dillingham, 2006)</p> <p>(Prasada, 2017)</p> <p>(Miller, 2019)</p>	<p>Inherent and extrinsic features</p>	<ul style="list-style-type: none"> - A unified cognitive process where recognising and explaining an object's characteristics based on its classification, distinguishes between inherent k-properties tied to the nature of the kind, and extrinsic t-properties, enhancing the relevance and effectiveness of explanations in abductive reasoning.
<p>(Lipton, 1990)</p> <p>(D. J. Hilton & Slugoski, 1986)</p> <p>(D. Hilton et al., 2005)</p> <p>(Lombrozo, 2010)</p> <p>(Woodward, 2006)</p> <p>(Miller, 2019)</p> <p>(Leddo et al., 1984)</p> <p>(McClure & Hilton, 1997)</p>	<p>Explanation Selection</p>	<ul style="list-style-type: none"> - Contrastive Explanations: People often explain events by contrasting them with a potential alternative scenario, emphasising the key aspects by comparison. - Abnormal Events: Highlight conditions considered abnormal to clarify why certain events occur outside the expected norm. - Intention and Function: Explanations that incorporate the intention behind actions or the function of objects are more satisfactory, attributing a deeper meaning. - Necessity, Sufficiency, and Robustness: These criteria are used to evaluate the strength of an explanation, where comprehensive explanations that require no additional factors are preferred. - Responsibility: Events deemed more responsible for an outcome are considered better explanations, closely linked to their necessity. - Preconditions, Failure, and Intentions: These elements add depth to explanations, providing context or motives behind actions.

<p>(Josephson & Josephson, 1996)</p> <p>(McClure, 2002)</p> <p>(Miller, 2019)</p> <p>(Thagard, 1978)</p> <p>(Read & Marcus-Newhall, 1993)</p>	<p>Explanation Evaluation</p>	<ul style="list-style-type: none"> - Explanations based only on statistical probabilities often feel insufficient because they lack a causal connection, leading to dissatisfaction as people prefer causality over mere statistical relationships. - Coherence, Simplicity, and Generality: Explanations are more effective when they are coherent with existing beliefs, simple, and broad, encompassing general principles rather than numerous specific details.
<p>(Chromik & Butz, 2021)</p> <p>(Miller, 2019)</p> <p>(Madumal et al., 2018)</p> <p>(Jin et al., 2021)</p>	<p>Multiple ways of explanation</p>	<ul style="list-style-type: none"> - Integrating multiple explanation methods addresses individual differences in information processing, enhancing understanding by allowing user interaction and exploration of alternative scenarios, especially in understanding causal relationships. - Utilising varied modalities like visual, textual, or auditory explanations caters to the specific context and audience background, improving engagement and trust through appropriately tailored communication. - Designing context-sensitive XUIs that adapt explanations to the domain knowledge and expertise of the user is crucial for effective model interpretation.
<p>(Miller, 2019)</p> <p>(Lombrozo, 2007)</p>	<p>Context dependent</p>	<ul style="list-style-type: none"> - Adjusting explanations in machine learning applications is essential to meet the varied expertise and experience levels of end-users, requiring simplicity for non-experts and detailed nuances for domain experts. - Different types of explanations should be tailored to the specific needs and familiarity levels of users, such as using heat maps for visualising model gradients in a way that matches the data type and user understanding. - Crafting effective explanations involves answering crucial questions about what to explain (identifying relevant data or features), how to explain (selecting suitable methods or tools), and whom to explain to (customising explanations for the audience's knowledge level and expectations).

3.1.11. Conclusion

This conclusion addresses the sub-research question:

"What are the factors that make an explanation interpretable for end-users?"

It becomes evident that interpretability hinges significantly on the human-centric nature of explanations. Effective explanations must resonate with end-users on a personal and intuitive level, closely aligning with natural human reasoning processes and social interactions. One of the primary insights from the review is the importance of explanations mirroring human social interactions, such as the expectation that artificial agents provide explanations in a human-like manner.

Moreover, the adaptability of explanations to the audience's biases, prior knowledge, and contextual understanding is crucial. While biases and prior knowledge can provide a scaffold that aids in understanding by aligning with pre-existing beliefs, they can also impede understanding when they lead to misconceptions or when the explanations contradict these beliefs. Managing the balance of information—avoiding overload while ensuring adequacy—tailors the content to the user's level of familiarity and expertise, thereby maximising interpretability.

Central to human understanding are the concepts of causality and contrastive reasoning. People inherently seek to understand cause-and-effect relationships, which are often best illustrated through scenarios that contrast what happened with what could have happened. This method simplifies complex information, making it more digestible and meaningful to the user. Additionally, abductive reasoning and counterfactual thinking are highlighted as valuable strategies for making explanations more plausible and comprehensible, particularly in complex systems where definitive answers may not be readily available. These reasoning strategies help users explore alternative scenarios and infer the most likely explanations, enhancing their understanding of causal relationships.

Lastly, the customisation of explanations to meet the specific needs and contexts of different audiences is paramount. This involves a careful consideration of what to explain, how to explain it, and to whom the explanation is directed. Adapting explanations in this way ensures they are not only understood but also appreciated by end-users, thereby fostering trust and confidence in AI systems.

In conclusion, achieving interpretability in explanations requires a deep understanding of human cognitive processes and social dynamics, as well as an ability to effectively communicate complex information. By focusing on human-centric design principles and tailoring explanations to the specific needs of users, AI systems can deliver explanations that are not only technically accurate but also meaningful and accessible to the people who use them.

3.2. Interaction

This section delves into the role of interactivity in enhancing the interpretability of end-users within XAI. It delves into the existing literature to examine and elaborate on various implementations of XUIs, explanatory goals, interaction types and key design principles that influence the development of effective and user-centric XUIs, and how these factors could collectively enhance user interpretability.

3.2.1. Explainable User Interfaces types

The development of Explainable User Interfaces (XUIs) has emerged as a pivotal area of focus within the fields of XAI and HCI, addressing the need to make underlying computational processes comprehensible to users. Various types of XUIs have been proposed by different authors, each designed to cater to specific user interactions and understanding levels.

DARPA's two-staged approach

The DARPA XAI program adopts a two-staged approach to developing explainable artificial intelligence, as outlined by Gunning and Aha (2019). This process distinctively separates the creation of the AI models and the interfaces through which users interact with these models.

The first stage is centered on the development of interpretable models that are inherently transparent. The aim here is to construct AI systems whose operational mechanisms are clear and comprehensible to users. This transparency is vital for ensuring that users can understand how decisions or predictions are made, which is crucial for building trust and facilitating effective management of the AI system.

The second stage focuses on the design of user interfaces that enhance how humans interact with the AI. This involves crafting tools for visualisation and interactive explanation platforms that enable users to query the AI and receive intelligible responses. The objective of this stage is to bridge the cognitive gap between the complex operations of AI models and user understanding, thereby enhancing the efficacy of human-AI collaboration.

By integrating these two stages—the development of transparent models and the creation of effective user interfaces—the DARPA XAI program aims to produce AI systems that are not only advanced in capabilities but also high in usability and trustworthiness.

Explanatory vs. Exploratory

Shneiderman (2020) categorises XUIs into two primary modes: explanatory and exploratory. Explanatory XUIs are structured to provide specific, focused explanations through visual or textual outputs. They are designed to clarify distinct elements of a model's functioning, making them ideal for presenting straightforward insights into particular aspects of the model. This mode is beneficial when users need direct and precise explanations without the necessity for deeper interaction with the model.

In contrast, exploratory XUIs offer a more dynamic interaction, allowing users the autonomy to probe and manipulate the model's behavior. This mode is incredibly advantageous when the interface permits users to alter inputs to see various outcomes, thereby fostering a more profound comprehension through hands-on engagement. Such interfaces not only enhance understanding but also increase user trust by transparently revealing how different inputs affect predictions.

An illustrative example of an exploratory XUI can be seen in a model predicting the probability of surviving the Titanic (Dijk, n.d.). This type of XUI would enable users to delve into detailed aspects of the model such as feature importances, SHAP values (feature contributions), what-if scenarios, dependencies, feature interactions, and even individual predictions and decision trees. For instance, users could examine how individual factors, like the deck location (lower or higher), passenger's sex, embarkation port, or fare paid, influenced survival probabilities. By interacting with these elements, users can gain insights into the predictive mechanics and even test hypotheses by modifying these factors to see how the survival predictions change.

Static vs. interactive explanations

Arya et al. (2019) differentiate between static and interactive explanations within XUIs. Static explanations are fixed and do not alter in response to user interactions, providing a consistent but limited understanding. Conversely, interactive explanations adapt to user queries and interactions, allowing for a more tailored and in-depth exploration of the model's behavior. This dynamic nature supports a

more comprehensive and user-centered explanation experience, as users can drill down into the details or shift the focus of explanations to suit their needs.

The design and implementation of XUIs require careful consideration of the type of explanations provided. It is crucial to balance between explanatory and exploratory elements based on the intended user base and the complexity of the model being explained. Additionally, the degree of interactivity and adaptability of the UI should be aligned with user feedback and needs, ensuring that the interface remains user-friendly while providing meaningful insights into the AI's decision-making processes.

Overall, the successful deployment of XUIs depends on a nuanced understanding of the different types of interfaces available and their appropriateness for various user scenarios. By integrating the right mix of explanatory and exploratory features, along with varying levels of interactivity, XUIs can significantly enhance user trust and comprehension of complex AI systems.

3.2.2. Explanatory goals in XAI

Tintarev (2007) identified seven key explanatory goals that can also be applied in the context of XAI design: transparency, scrutability, trustworthiness, persuasiveness, effectiveness, efficiency, and satisfaction. Each of these goals serves a distinct purpose in the design and functionality of user interfaces, aiming to bridge the gap between AI operations and user understanding.

1. **Transparency** This goal focuses on clarifying how the AI system functions. By illuminating the inner workings of the AI model, transparency helps users understand the basis on which the system makes decisions or predictions.
2. **Scrutability** It is crucial that users have the ability to scrutinise, question, and if necessary, correct the system. This fosters a deeper understanding and control over AI interactions, enhancing user engagement and trust.
3. **Trustworthiness** A key objective is to build a system that users can trust. Trustworthiness in AI systems assures users of the reliability and integrity of the explanations and decisions provided.
4. **Persuasiveness** The system should deliver explanations that are not only clear but also compelling enough to convince users of their validity. Persuasive explanations can lead to greater acceptance and reliance on the system's recommendations.
5. **Effectiveness** Effective AI systems aid users in making better-informed and more accurate decisions, thereby improving the overall decision-making process.
6. **Efficiency** Explanations should be designed to help users reach decisions quickly and with less effort, streamlining user interactions with the system.
7. **Satisfaction** Ultimately, the interface should offer a satisfying and positive user experience, making interactions with the AI system enjoyable and rewarding.

However, achieving a balance among these goals can be challenging as potential trade-offs may exist due to conflicting priorities (Tsai & Brusilovsky, 2019). For instance, in the context of a specific use case such as the FOKUS project, priorities shift towards persuasiveness, effectiveness, and efficiency, possibly at the expense of other goals like transparency or scrutability. This is particularly relevant in scenarios where rapid decision-making and user conviction are critical.

The emphasis on different goals often depends heavily on the application context and the specific needs of the users involved. Pursuing one goal more aggressively could inadvertently diminish the impact of another, necessitating a careful consideration of how these goals are prioritised and balanced in the design of XAI systems. This nuanced approach ensures that XAI systems are not only technically proficient but also aligned with the diverse expectations and requirements of their end-users.

3.2.3. Interaction defined

Interaction has been defined in various ways in the literature. Miller (2019) characterises XAI as a type of human-agent interaction problem, where an explanatory agent reveals the underlying causes behind its own or another agent's decision-making process. In this context, the interaction occurs between a human user and an AI agent, facilitated by an XUI.

Hornbæk and Oulasvirta (2017) describe interaction as the interplay between two or more constructs. They conducted an analysis of the interplay between the human and computer constructs, which has been widely discussed in the field of Human-Computer Interaction (HCI) research. Based on their analysis, they derived seven concepts of interaction: information transmission, dialogue, control, experience, optimal behavior, tool use, and embodied action.

Building on these concepts, Chromik and Butz (2021) applied them to the context of human-XAI interaction. By leveraging these different perspectives, they explored and examined how the human and AI agent interact within an XAI system.

- **Interaction as (Information) Transmission.** Most basic form without offering any other explanations, the XUI is only used as a tool to display the outputs. The main objective of this interaction is to provide users with a comprehensive explanation. Various publications focused on this concept emphasise the importance of transparency and acknowledge that algorithms should not be studied in isolation. Instead, they should be examined alongside interfaces, as both components significantly contribute to the perception of explainability. The goal is to understand and present a complete explanation that considers both the inner workings of the algorithm and the user interface.
- **Interaction as Dialogue.** A cycle of communication of inputs/outputs by the computer and perception/action by a human. Happens in stages or in turn, where the user asks a question and the AI answers the question. There is a distinction made between the type of dialogue.
Inspection Examples : Exploratory dialogues provide users with the opportunity to investigate how potential changes in inputs can influence the AI's predictions. It allows users to examine the inner workings of the AI system. The XUI primarily focuses on providing functionalities that enable users to repeatedly request explanations of the same nature. This iterative process allows users to gain a deeper understanding of the AI system and its predictions.
Natural examples : have the objective of reducing the level of expertise required to analyse data, thereby enhancing the accessibility of XUIs for end users of XAI. The focus of XUIs is to provide users with functionalities that allow them to request various explanations using natural language. In this interaction, the human user takes the lead by asking questions, and the XUI responds with explicit explanations in the form of simplified natural language textual answers.
- **Interaction as control :** This concept facilitates the efficient and consistent progression of the human-computer system towards a desired state. Drawing from control theory principles, the interaction aims to adjust a control signal to reach a specific level and continuously adapt its behavior based on feedback.
- **Interaction as Experience.** Human expectations play a crucial role in the interaction with computers. By effectively managing these expectations, even if the accuracy of the system is not at its highest level, users can still find the system useful as long as it remains transparent about its limitations. When users are aware of the system's capabilities and potential shortcomings, they can adjust their expectations accordingly and appreciate the system for what it can deliver. The literature indicates that insufficient mental models and a lack of understanding about the system can result in failed interactions (Bethel, 2009; Chandrasekaran et al., 2017). Transparency about accuracy fosters trust and allows users to make informed decisions about utilising the system's capabilities effectively.
- **Interaction as Optimal Behavior.** While this point is closely related to the previous point, this one focuses more on altering the behavior of the user and how they navigate through the UI instead of what they expect. Difference in this can be made between i. the limitations that occur during interaction and ii. design interaction to better accommodate the limitations.
i. Examine limitations : To examine the interactions, the cognition preferences of users need to be taken into account during their interpretation of the explanations. E.g. their tendency to engage, their first impression, and to direct user's attention to the strong points of the system as opposed to the weaker ones.
ii. Moderate limitations : How to moderate the limitations. Provide explanations as to why the human should trust the AI or not. Highlighting the ambiguous predictions of the model to offload their cognitive resources so they can judge themselves how the model should be interpreted. This can also be done by using visual explanations which balancing the cognitive loads by limiting the vi-

sual chunks. Important is to let the user also know why the AI has limitations and let them know what the limitations are.

- **Interaction as tool use.** This interaction concept assists humans in uncovering hidden patterns and insights within domain-specific data. To support this learning process, some form of explanation is necessary. The XUI acts as a tool that provides a unique perspective on a particular domain, going beyond just the behavior of the AI system. It enables users to comprehend complex information that would otherwise be challenging to understand.
- **Interaction as embodied action.** Both parties actively adjust their expectations and continuously communicate their capabilities in real-time to achieve a shared objective. The XUI not only provides information and explanations but also influences the user's actions. Similarly, the user's actions and inputs can influence the behavior and responses of the XUI. This interaction is a dynamic process driven by both the XUI and the user, with constant feedback and mutual influence, working together towards a common goal.

3.2.4. Design Principles

Chromik and Butz (2021) has effectively refined the varied interactions between humans and AI systems explained in the previous section into a set of design principles tailored specifically for interactive XUIs. There are four key design principles proposed by Chromik and Butz that will be pivotal in shaping the XUI design: Complementary Naturalness, Responsiveness through Progressive Disclosure, Flexibility through Multiple Ways to Explain, and Sensitivity to Context and Mind. Each of these principles offers a unique angle on enhancing user experience and interpretability, aligning closely with the diverse needs and cognitive processes of users. These principles will be explored in further detail in the subsequent discussion.

1. Complementary Naturalness

This principle refers to the idea of combining implicit visual explanations with natural language rationales. Implicit visual explanations, while accurately depicting the inner workings of an AI, can often be challenging for non-experts to comprehend. On the other hand, natural language rationales are post-hoc explanations designed to mimic what a human explainer would say in a similar situation. Textual rationales can provide reassurance to users when the explanations provided by the system are uncertain or unclear. By combining visual explanations with textual rationales, both understanding and communicative effectiveness can be enhanced, making the overall explanation more accessible and informative for users.

2. Responsiveness through progressive disclosure

This principle refers to the use of hierarchical or iterative functionalities that enable follow-up explanations based on initial explanations. The threshold for the amount of explanation needed varies for each user, based on their individual need for understanding. Some users may have a simpler mental model of the underlying AI and may be overwhelmed by overly detailed explanations. It recognises the variability in users' desire for and ability to understand complex information. By gradually disclosing additional information and tailoring the level of detail to the user's comprehension, the system can strike a balance and provide explanations that are informative without overwhelming the user.

3. Flexibility through multiple ways to explain

Acknowledging the diversity in human understanding and preferences, this principle advocates for multiple explanatory methods. In practical terms, there is often no single "best" way to explain something. For example, when a physician is making a differential diagnosis, they consider multiple types of data and information. Similarly, in the context of explanations, different methods and modalities can complement each other. By utilising a combination of explanation approaches, such as visual, textual, or interactive, the system can cater to different learning styles and provide a more comprehensive understanding of the subject matter. This flexibility allows for a more effective and adaptable explanation process.

4. Sensitivity to the mind and context

This principle underscores the importance of customising explanations to fit individual user characteristics. As users interact with the system, their understanding deepens and their needs for explanations may evolve, necessitating adjustments in the information provided. Moreover, personal biases and

prior beliefs significantly shape how users perceive and react to explanations. An effective XAI system, therefore, must employ a personalised approach that continually adapts to these changing needs. This involves considering each user's unique cognitive processes, preferences, and contextual factors to craft explanations that are both relevant and impactful. Through such tailored interactions, the system not only enhances the interpretability of the AI but also fosters a deeper, more meaningful engagement with the user.

3.2.5. Conclusion

This conclusion addresses the sub-research question:

"How can interaction influence the interpretability of end-users?"

Interactivity in XAI primarily enhances interpretability through two main avenues: increasing transparency and fostering user engagement. Different types of interactions—such as information transmission, dialogue, and control—each contribute uniquely to this process. Information transmission, the most basic form of interaction, ensures that users receive straightforward outputs from AI, while dialogue and control allow for a more dynamic exchange where users can query the AI and even influence its operations.

The development of transparent AI models, exemplified by DARPA's two-staged approach, creates a base level of user understanding by making the operational mechanisms of AI systems clear. This transparency is foundational for building trust but becomes truly effective when coupled with interactive Explainable User Interfaces (XUIs). These interfaces, ranging from static to interactive, allow users to actively manipulate data, explore what-if scenarios, and see the direct impact of changes in inputs on outputs. Dynamic interactivity such as this turns static information into a collaborative exploration, demystifying AI decision-making processes and significantly boosting interpretability.

Moreover, the interactivity facilitated by XUIs encourages users to move from passive reception of information to active participation. Interfaces that support extensive dialogue and offer control enable users to question, challenge, and modify AI functionality. This not only deepens understanding but empowers users to trust and rely on the technology. By transforming interpretive processes from a monologue into a dialogue, these interfaces make AI systems more accessible and comprehensible.

The design principles proposed by Chromik and Butz—Complementary Naturalness, Responsiveness through Progressive Disclosure, Flexibility through Multiple Ways to Explain, and Sensitivity to Context and Mind—serve as critical guidelines for enhancing the user experience and interpretability. By combining visual and natural language explanations, XUIs become more accessible and informative, making complex AI operations understandable to non-experts. Progressive disclosure adapts the depth of information to user comprehension levels, preventing overload and enhancing engagement through iteratively deepened explanations. Multiple explanatory methods address the diversity in user preferences and learning styles, offering a richer, more adaptable interaction. Lastly, sensitivity to the individual user's context and cognitive processes ensures that explanations are not only relevant but also resonate on a personal level, fostering deeper understanding and trust.

However, balancing the explanatory goals in XAI—such as transparency, scrutability, trustworthiness, and efficiency—presents challenges due to inherent trade-offs. For instance, increasing transparency might reduce efficiency, and enhancing persuasiveness might at times limit scrutability. The literature suggests that achieving an ideal balance among these goals is complex and dependent on the specific contexts and needs of the users.

In conclusion, the types and depth of interactivity in XAI are powerful drivers for enhancing user interpretability, effectively addressing the sub-research question. By leveraging varied interaction types—from straightforward information transmission to complex, controllable dialogues—XAI can meet diverse user requirements. This not only helps in making AI decisions transparent and trustworthy but also ensures that users can effectively leverage (X)AI insights in their decision-making processes.

4

Design

This chapter introduces the use case provided by FOKUS, which serves as the basis for this research. It outlines the background, highlights key concerns, and acknowledges the limitations inherent in the use case. Subsequently, the chapter aligns the findings from the literature review with the specifics of the use case to construct both a conceptual and a theoretical framework. Building on this foundation, the design of the XUI will be developed by applying the applicable interactive design principles for this use case by Chromik and Butz (2021).

4.1. Use Case: FOKUS

The deliverable of this research is a preliminary prototype designed which could be used for future development. The prototype is founded upon a practical use case provided by Fraunhofer FOKUS, an institute specialising in open communication systems (FOKUS, 20223). One of the notable projects currently underway at FOKUS involves the research on a wearable emergency medical device aimed at monitoring vital signs, specifically through Electrical Cardiograms (ECG). The project involves the application of an AI which detects symptoms indicative of a Myocardial Infarction (MI), commonly known as a heart attack, in real-time. The project serves as a use case for FOKUS for the purpose of demonstrating an approach that could be taken for an explainable machine learning model.

The AI system's functionality includes the detection of anomalies in ECG readings and the activation of an alert mechanism to promptly notify emergency services in the case of a heart attack. This feature is particularly crucial as the timely detection and response can be a matter of life and death. A vital step in this process involves the AI's preliminary findings being evaluated by a human expert—typically a healthcare professional with expertise in reading ECGs and diagnosing heart conditions—before any definitive action is initiated. This hybrid decision-making model, where AI speed and accuracy are complemented by human expertise, ensures both swift and reliable responses in critical medical situations.

To facilitate this critical human-AI interaction, FOKUS is also actively developing an XAI approach tailored to articulate the AI's decision-making process to the human evaluator. This development involves researchers at FOKUS who are exploring various XAI methodologies to determine the most effective ways to present complex AI analyses in an understandable manner.

The three XAI methods under consideration at FOKUS include SHAP (SHapley Additive exPlanations), Grad-CAM (Gradient-weighted Class Activation Mapping), and LIME (Local Interpretable Model-agnostic Explanations). While the operational mechanics of each method are complex, a brief overview of the methods used and the dataset employed for the AI is provided in section 4.1.2. This overview illustrates how these methods enhance the interpretability of AI decisions within the healthcare context. The research initiative at FOKUS is not only aimed at evaluating these methods for their effectiveness but also at integrating the most apt technique into a sophisticated XUI. Currently, an effective XUI has not been developed, but is crucial for the next steps. Findings in the literature review suggest that the XAI is challenging to comprehend without an effective interface, underscoring the need to develop an

XUI that facilitates this crucial intermediate step. The forthcoming XUI will be specifically tailored to this case and the corresponding XAI, enabling human experts to conduct precise and efficient assessments based on the AI's analyses, thereby ensuring a rapid and reliable medical response in emergency situations.

4.1.1. ECG data

ECG data captures the electrical activity of the heart over a period, represented as line tracings on paper or screens. These tracings, composed of waves and segments, provide critical information about the heart's rhythm and electrical activity. Myocardial Infarction (MI), commonly known as a heart attack, is identified in ECGs through specific patterns. These include changes in the ST segment, T wave inversions, and the presence of new Q waves. Such patterns reflect disturbances in heart muscle activity due to insufficient blood supply, which is indicative of an MI. By analysing these elements in the ECG waveform it can be determined whether a person is experiencing or has experienced a heart attack (Hampton & Hampton, 2019). An overview of the components of the ECG complex can be seen in Figure 4.1

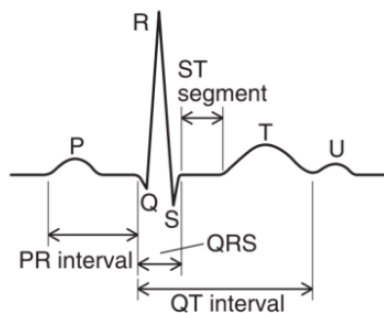


Figure 4.1: Components of the ECG complex (Hampton & Hampton, 2019)

ECG data is typically displayed on a grid layout, with each grid cell measuring the heart's electrical activity in a structured format. The standard display comprises a 12-lead ECG, which records the heart's electrical signals from 12 different perspectives by using electrodes placed on the patient's limbs and chest. These leads include three limb leads (I, II, and III), three augmented limb leads (aVR, aVL, and aVF), and six chest leads (V1, V2, V3, V4, V5, and V6). This layout is often presented in a 4x3 format on screens or paper (Figure 4.2), where each row represents a different set of leads grouped by similarity in viewpoint or anatomical position. This configuration helps in systematically analysing the heart's electrical function and identifying abnormalities (Hampton & Hampton, 2019).

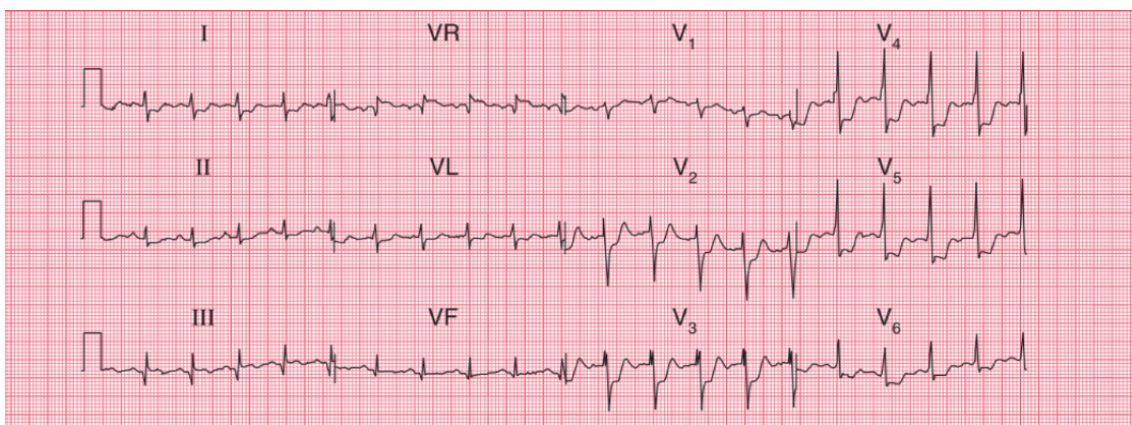


Figure 4.2: Standard 12 lead ECG display (Hampton & Hampton, 2019)

4.1.2. FOKUS XAI

XAI researcher perspective

At FOKUS, researchers conducted an extensive literature review on the interpretation of ECGs to guide the development of their (X)AI model. This review highlighted specific ECG leads as crucial for detecting signs of Myocardial Infarction (MI), which is a primary focus of their study. This information has been integral to training their models. It's important to note that the researchers are not aiming for perfect explanations but rather seek to demonstrate an effective method for developing explainable models. As their work progresses without direct input from medical experts, they rely predominantly on documented research and existing knowledge, focusing on exploring XAI design methodologies rather than creating clinical products.

For visual explanations, the researchers have based their choices on both literature insights and intuitive design considerations. The layout of the ECG display was selected for its usability across different XAI methods, facilitating comparisons and assessments. The use of colour in visual explanations is experimental, varying with the type of data shown; for instance, positive or negative indications might use distinct colours or a gradient scale to denote varying degrees of impact or severity.

The researchers are confident that their XAI outputs contain all necessary elements for a medical expert to diagnose MI effectively. However, they recognise challenges with the XAI quality, attributable to the project's nascent stage. Presently, significant issues include inconsistencies across various XAI methods and within the same method when processing identical datasets. The focus is predominantly on rectifying these inconsistencies rather than on refining the visual presentation.

AI dataset

To train their AI for the detection of MI, the researchers at FOKUS have used the PTB-XL dataset, a large publicly available electrocardiography dataset, which has been put forward and discussed by Wagner et al. (2020). The PTB-XL dataset is the largest freely available clinical 12-lead ECG-waveform dataset, containing 21,837 records from 18,885 patients, each 10 seconds long. It features multi-label annotations by cardiologists, organised into diagnostic categories including a substantial number of healthy records. Enhanced with detailed metadata such as demographics, diagnostic statements, and manually annotated signal properties, PTB-XL is a valuable resource for developing and evaluating ECG interpretation algorithms. It addresses key challenges like the scarcity of public datasets and standardising benchmarking procedures for algorithm comparison. The dataset encompasses a comprehensive range of diagnostic classes, a detail that will be crucial to the findings presented later in this study.

XAI outputs

Included below are examples of the current XAI outputs being developed at FOKUS. These images illustrate the researchers' methodological choices and the integration of their scholarly findings into the XAI outputs. The layout of the ECG traces are presented with all the 12 leads stacked on top of each other. The ECG traces are displayed with all 12 leads stacked vertically, each lead marked on the left side with its corresponding voltage division. To the right, a colour bar correlates numbers to a colour gradient, representing the impact score assigned by the XAI to specific points on the AI model's output. Time divisions are marked at the bottom in ms/V, showcasing 2.5 seconds (250ms) of ECG data. At the top, details such as the patient ID, the true diagnosis of MI (true), the AI diagnosis score of MI (predicted), and the method used are displayed. The layout of additional elements surrounding the central XAI output varies with the method employed, details of which are provided in subsequent sections. The underlying AI models will not be discussed as this falls outside the scope of the thesis.

SHAP

SHAP (SHapley Additive exPlanations) is an XAI method that offers explanations for models handling time series data. It employs a unique approach by attributing significance to each time point or feature in relation to the model's predictions. SHAP calculates Shapley values for all possible combinations of features or time points, quantifying their importance based on their impact on the model's output. This method effectively highlights which specific time points or features play a pivotal role in influencing the model's decision-making process (Lundberg & Lee, 2017). This method incorporates negative impact points, employing a colour gradient ranging from red to blue to distinctly highlight the differences in impact. The output for this method is shown in Figure 4.3.

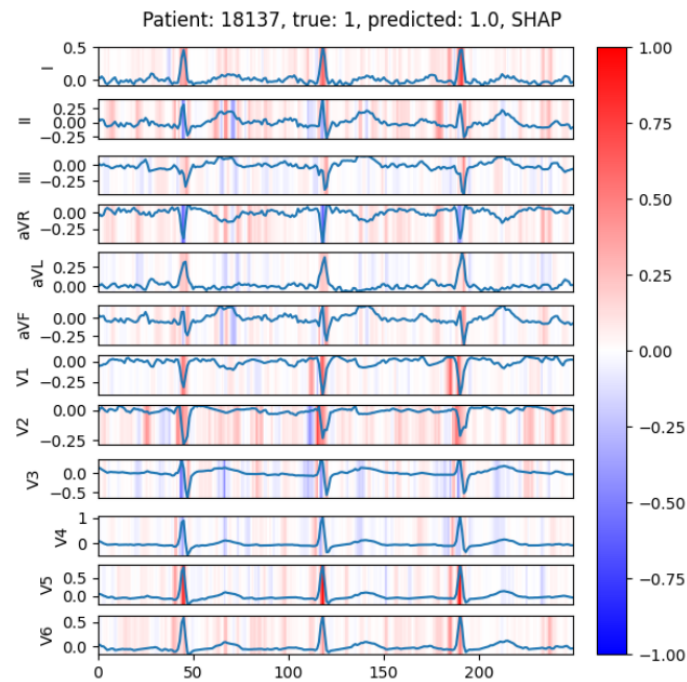


Figure 4.3: SHAP with positive and negative indications

LIME

LIME (Local Interpretable Model-agnostic Explanations) is an XAI method that explain how AI models analyse images by using segmentation in images. It strategically modifies images—adjusting attributes like sharpness and color saturation—to pinpoint which features crucially influence the model's predictions. Similar to Grad-CAM, the process may result in blurred or altered images as a by product of the computational method used, revealing how different image aspects impact the AI's decision-making (Ribeiro et al., 2016). This method highlights areas of importance without incorporating negative impacts, using a multi-coloured gradient to denote the varying significance of each area's impact on the prediction. The output for this method is shown in Figure 4.4.

pred label: norm, pred score: 0.65347, true label: mi, patient id: 7990

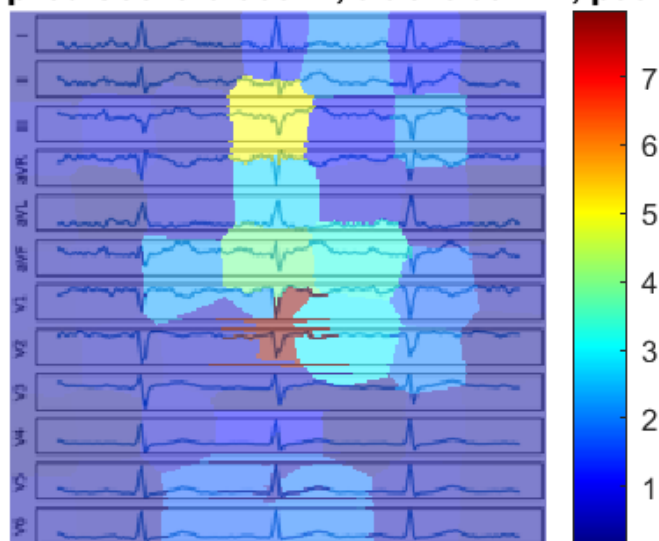
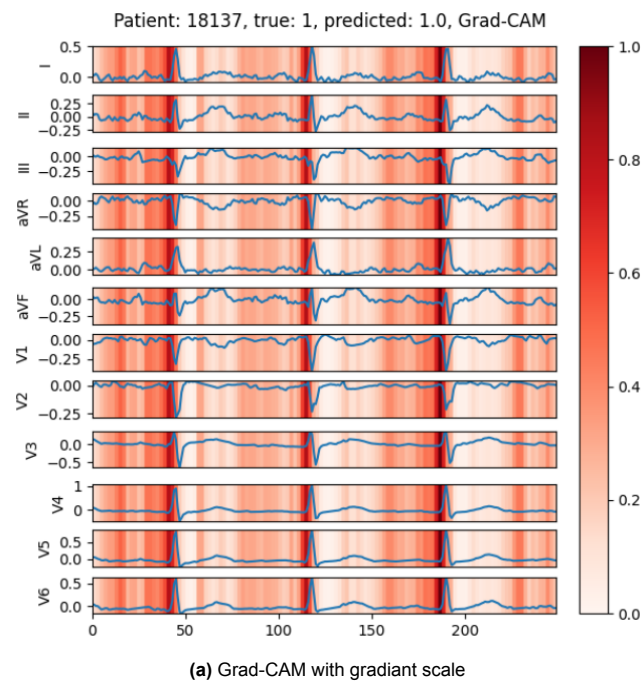


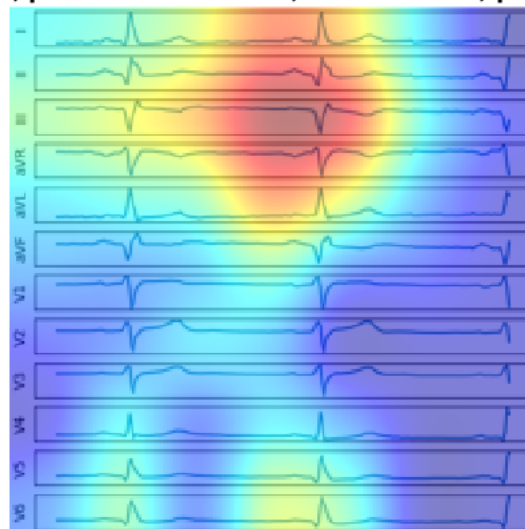
Figure 4.4: LIME with varying colour scale

Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is an XAI method that provides explanations for AI models working with images. Grad-CAM focuses on understanding which parts of the image the AI model "looks" at while making its predictions. It does this by analysing the model's gradients, which represent how much each pixel in the image contributes to the final prediction. Grad-CAM then generates a heatmap that highlights the most relevant regions of the image that influenced the AI's decision. It's possible that the images become blurry or altered due to the computational process, which allows Grad-CAM to efficiently explore various possibilities and provide meaningful explanations without overwhelming computing power requirements (Selvaraju et al., 2019). Two different outputs are currently being explored at FOKUS depicted in Figure 4.5a for the first output and Figure 4.5b for the second. Output 1 employs a red colour gradient and is based on a different underlying AI model, contributing to its distinct appearance compared to Output 2, which uses a multi-coloured gradient to represent the data.



pred label:mi, pred score:0.94848, true label:mi, patient id:15813



(b) Grad-CAM with varying colour scale

Figure 4.5: Grad-CAM outputs with different colour gradients

4.1.3. Concerns and limitations

- A concern emerged during the analysis of the use case regarding the display format of the ECG leads by FOKUS, which deviates from the standard format typically illustrated in academic and clinical settings, as shown in Figure 4.2. Upon raising this issue, FOKUS responded that they had not prioritised the standard layout in their visual explanations, and did not consider that the layout could be part of the explanation interpretation within the scope of their research objectives. However, they acknowledged the potential relevance of adhering to conventional display norms and expressed openness to revisiting this aspect in future developments, though adjustments could not be accommodated within the current thesis timeline due to time constraints.
- Limitations emerged in the types of explanations that could be generated within the XAI methods used, primarily influenced by the specific nature of the XAI methods and the developmental phase at FOKUS. Several enhancements were contemplated to increase the interpretability of the XAI outputs; however, these enhancements were not feasible due to existing constraints. Considered additions included the integration of counterfactual scenarios, such as displaying cases of patients with similar symptoms but without an MI diagnosis, to provide comparative insights. Moreover, improvements such as incorporating higher-resolution images for certain XAI methods and enabling zoom functionality on ECG traces for detailed examination were also evaluated but could not be implemented at this stage of development. Design elements and considerations will be elaborated in section 4.3.
- A misunderstanding arose regarding the expectations from both the research team at FOKUS and the thesis researcher. FOKUS was primarily focused on developing an explainable AI model rather than refining the explanations themselves, which explains why they had not yet consulted with ECG experts. However, they were interested in engaging with medical experts to enhance their understanding of how specialists interpret ECGs to identify MIs, aiming to refine their models accordingly. Meanwhile, the thesis researcher planned to conduct interviews with field experts to assess the effectiveness of the designed XUI. The FOKUS team presumed these interviews might also provide insights for ECG interpretation useful for their model development. However, the thesis researcher clarified that the interviews would solely collect feedback on the XUI and would not delve into specific ECG interpretation or diagnostic questions. This misalignment of expectations initially caused some confusion. Once clarified, the collaboration proceeded with a realigned focus solely on the development and assessment of the XUI, ensuring that both parties were working towards complementary but distinct objectives.

4.1.4. Selection of use case

The case study highlighted in this analysis revolves around an XAI project currently being developed at FOKUS. Despite being in its nascent stages, this particular project was chosen for its strong potential to substantially enhance the research being conducted. It offers a distinctive view into the ongoing developments and the complex challenges faced within the XAI sector. Serving as a dynamic and active example, this project sheds light on the practical difficulties and obstacles encountered in real-world applications, making it an ideal candidate for a deep dive into the subtleties of enhancing interpretability in XAI.

Furthermore, this case study could stand to reveal unique challenges and critical considerations that might not yet be fully acknowledged or understood within the broader scientific community. By examining this specific XAI initiative at FOKUS, there is an opportunity to identify new problems and generate insights that could benefit a broader scope beyond this XAI initiative.

Therefore, the investigation into the XAI project at FOKUS is not merely about adding another example to the academic literature; it represents a deliberate and thoughtful approach to uncovering tangible, real-world challenges faced by researchers in the field. This exploration is intended to contribute to the wider conversation about the evolution of XAI technology, aiming to enrich both the theoretical framework and practical applications within the field. The ultimate goal is to advance the understanding of XAI development processes and to highlight novel challenges and solutions that can benefit the entire domain.

4.2. Aligning literature with use case

Building upon the use case outlined earlier, this section aims to align relevant literature with the practical aspects of the project to construct a robust theoretical and conceptual framework for this thesis. The literature review will integrate findings from previous studies and theoretical constructs to underpin the methodologies and approaches employed in the use case provided by FOKUS. This alignment will ensure that the research is grounded in established theories while also addressing the specific challenges and requirements of the XAI project.

4.2.1. Conceptual framework

The conceptual framework, illustrated in Figure 4.6, combines the findings of the literature review and practical application derived from the use case at hand. The literature suggests that incorporating interactivity into an XUI can increase an end-user's ability to comprehend an XAI system. However, this interpretability could be subject to the influences of individual biases and pre-existing knowledge, varying with the end-user's level of expertise and experience in both the domain context and with (X)AI technology.

The framework unfolds sequentially in three integral stages:

Firstly, it acknowledges the 'definition of context' as an ongoing process, one that evolves in tandem with the development of the XAI. This context is multifaceted, encompassing both the target user group and the domain-specific setting. Simultaneously, the characteristics and capabilities of the XAI itself are shaped by these same contextual elements. These foundational aspects dictate the feasible range of XUI design strategies, particularly the interactive features that can be feasibly integrated.

Secondly, within the framework of this thesis, the main hypothesis asserts that the implementation of interactive elements within the XUI is anticipated to increase interpretability. It further hypothesises that this potential increase is moderated by the user's ECG expertise and AI knowledge, suggesting that these factors may influence the extent to which interpretability is enhanced. The thesis contends that a well-crafted interactive interface, tailored to the user's context and expertise, can bridge the gap between complex AI outputs and user comprehension.

Lastly, the conceptual framework is not merely an academic exercise; it serves as a practical tool that will steer the subsequent design, validation, and conclusion chapters of this research. Through rigorous validation, the research will aim to substantiate the initial hypothesis and, in doing so, contribute meaningful insights to the body of knowledge surrounding interactive XUIs in enhancing the interpretability of XAI systems.

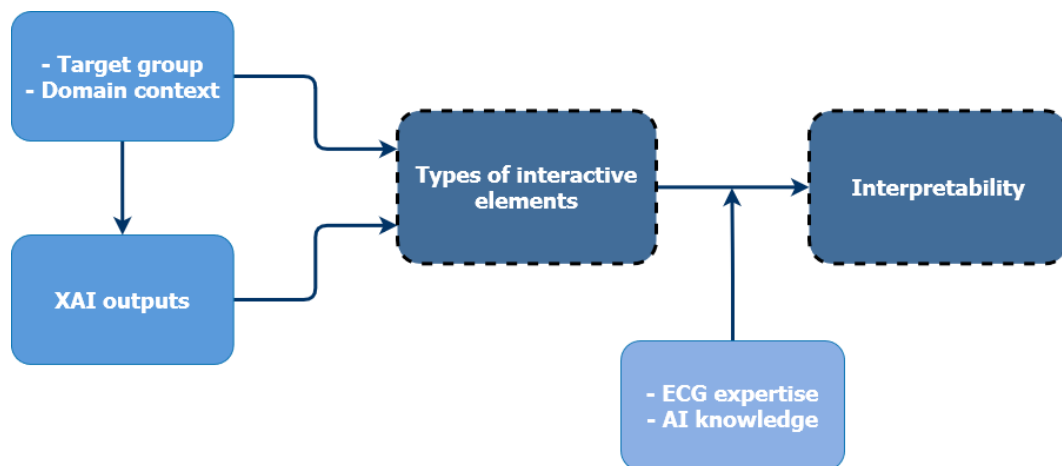


Figure 4.6: Conceptual framework

4.2.2. AI in the medical domain

To better understand the context of the use case, a concise literature review was conducted, focusing on the role of AI in the medical domain. This review provides insights into the current applications and challenges of AI in healthcare, offering relevant points that could inform the development of the XUI.

XAI in healthcare

The integration of Explainable Artificial Intelligence (XAI) in healthcare is pivotal for enhancing transparency and fostering trust between AI systems and medical professionals. As noted by A. and R. (2023), XAI's ability to clarify predictive modeling outcomes is crucial for its practical application in healthcare settings. Loh et al. (2022) further emphasise the confidence that XAI instills in AI predictions, encouraging its broader utilisation. Hulsén (2023) highlights the necessity of XAI in combating the "black box" nature of traditional AI methods, thus promoting understanding and trust in this essential area. Y.-C. Wang et al. (2023) demonstrate the application of XAI in making hospital recommendations, showcasing the cross-domain potential of these techniques.

In diagnostic imaging, particularly for breast cancer, AI exhibits remarkable capabilities in lesion identification and risk assessment

textbfcitelei2021. The development of visualisation systems further illustrates progress in imaging technologies, enhancing early detection efforts. XAI's role extends to explaining AI-driven decisions, crucial for diagnosing and planning treatments for breast cancer, and personalising medicine by elucidating the reasons behind specific treatment suggestions based on genetic profiles. This clarity is vital for clinicians to tailor effective interventions and manage disease progression expectations.

Additionally, in the realm of surgical robotics, XAI contributes to safety and efficacy by clarifying the rationale behind decisions made during procedures, thus complementing surgeons' expertise (O'Sullivan et al., 2022). In clinical trial matching, XAI aids both clinicians and patients by simplifying the decision-making process, clearly explaining why certain trials are recommended based on detailed patient profiles. Through these diverse applications, XAI not only enhances understanding but also fosters a collaborative approach to patient care, ultimately improving outcomes and building trust in AI technologies across the healthcare spectrum (Y.-C. Wang et al., 2023).

Ground truth in medicine

Having a reliable and accurate reference point, known as the ground truth, is crucial. The ultimate objective is for a statement or diagnosis to align perfectly with this ground truth (Ali et al., 2023), and any explanation provided for the statement should accurately correspond to this ground truth. However, in the medical domain, determining a definitive ground truth can be challenging. The most effective machine learning models in this field often rely on concepts such as correlation, similarity, and distance, rather than a clearly defined ground truth (Alfwzan et al., 2024).

Explanation in Medicine

What constitutes a good explanation can vary depending on the expertise of the individual receiving the explanation (Ali et al., 2023). For instance, a visual explanation that highlights the location of a disease might be satisfactory for a radiologist or a medical image analysis researcher (Müller et al., 2021). However, clinicians such as oncologists, neurologists, or hematologists would likely prefer to have XAI integrated into their clinical decision-making process (Antoniadi et al., 2021). Additional information that should also be taken into account are the patient's medical history, past and ongoing treatments, available treatment options, and the expected effects or outcomes of those treatments (van der Velden et al., 2022).

Scope of the XUI

In collaboration with researchers at FOKUS, the scope of the XUI was defined through a detailed exploration of the capabilities of the XAI methods and their potential translation into design elements. This iterative process involved using findings from the literature review in Chapter 3 and the previous section 4.2.2 about AI in the medical domain, to conceptualise possible interface elements through mind mapping. The focus was on implementing the interactive design principles outlined by Chromik and Butz (2021), while considering the types of interactivity and interpretability from the literature as supporting factors. The thesis researcher and the FOKUS team engaged in discussions to evaluate the feasibility of incorporating these elements into the XUI. To effectively define the scope, three critical

questions were addressed, identified as essential building blocks for crafting effective explanations in machine learning by Miller et al. (2017). These questions, as previously mentioned in the literature review under section 3.1.10, helped guide the development process and ensure the relevance and effectiveness of the XUI design.

- **What to explain?** This question is approached through two key contexts related to the use case. From FOKUS' side, the aim is to assist medical experts decision making to determine MIs using XAI methods. These methods are still in the early stages of development and currently provide static outputs. Extending this to align with the thesis, the secondary context is to integrate features in the XUI that explain how these XAI methods function with the aim to increase their interpretability. This is intended to help experts in their clinical decisions regarding MI diagnoses. Considering that this may be the first interaction for many experts with the XUI, and possibly XAI in general, it is essential that the XUI offers explanations that are clear and easy to understand, avoiding information overload. The overarching challenge lies in merging these dual objectives—enhancing understanding of the model and aiding in decision-making—within the XUI design, creating an interface that both educates and enables practical usage by healthcare professionals.
- **How to explain?** The primary mode of explanation provided by the XAI methods will be visual, directly derived from their outputs. Surrounding these primary visuals, the XUI will incorporate alternative forms of explanation to enrich the user's understanding. This may include textual interpretations of the visual data, supplementary patient information not directly displayed by the XAI methods, and details that might be challenging to discern visually by the human eye. Currently, the XAI systems do not support interactive features that allow for the modification of model outputs by the user; however, they can display additional context such as the most relevant ECG leads or prediction confidence scores.
- **Whom to explain to?** The primary target audience for this use case consists of medical professionals equipped to diagnose MIs from ECG data. To ensure a broader and more diverse participant base, this target group will be expanded to include professionals actively engaged in the medical field who possess a basic understanding of MIs and ECGs. It is anticipated that most members of this group will have limited familiarity with (X)AI, as their primary focus lies outside the realm of AI technology. This demographic consideration is crucial to tailor the explanation style and complexity in the XUI, ensuring it is accessible and valuable to those with little to no prior experience with (X)AI systems.

By addressing these pivotal questions, it was concluded that three of the four design principles outlined by Chromik and Butz (2021) could be integrated into the XUI development. These principles are complementary naturalness, responsiveness through progressive disclosure, and flexibility through multiple ways to explain. The principle of sensitivity to context and mind was omitted in this initial prototype of the XUI, primarily due to the inability of the current system to support the personalisation required to adapt explanations to each individual's unique context and cognitive preferences, which is central to this design principle.

With the parameters of the XUI now defined and the applicable design principles selected, the next steps involve assessing how the remaining interaction and interpretability factors from the literature can be integrated according to these principles. This analysis will be detailed in the subsequent section.

4.2.3. Theoretical Framework

Building on the extensive literature discussed in Chapter 3 and the previous section 4.2.2, a theoretical framework was established, synthesising key concepts relevant to the study. Given the specific focus of this thesis, the constraints of time and scope presented by the FOKUS use case, not all elements from the literature could be directly applied. The core objective is to implement specific design principles that underpin the theoretical framework, tailored to the use case provided by FOKUS.

The development of this framework commenced with the selection of appropriate design principles, as outlined in the preceding section 4.2.2. This was followed by identifying the types of interactivity suitable for the scope of this XUI, each linked to an appropriate design principle. Subsequently, the factors influencing interpretability were refined to align with the scope of the thesis and connected to the corresponding types of interactivity. This structured approach ensures that the XUI design is thoroughly grounded in theory.

Figure 4.7 illustrates how the chosen interpretability factors are associated with specific types of interactivity, which in turn are linked to the designated design principles.

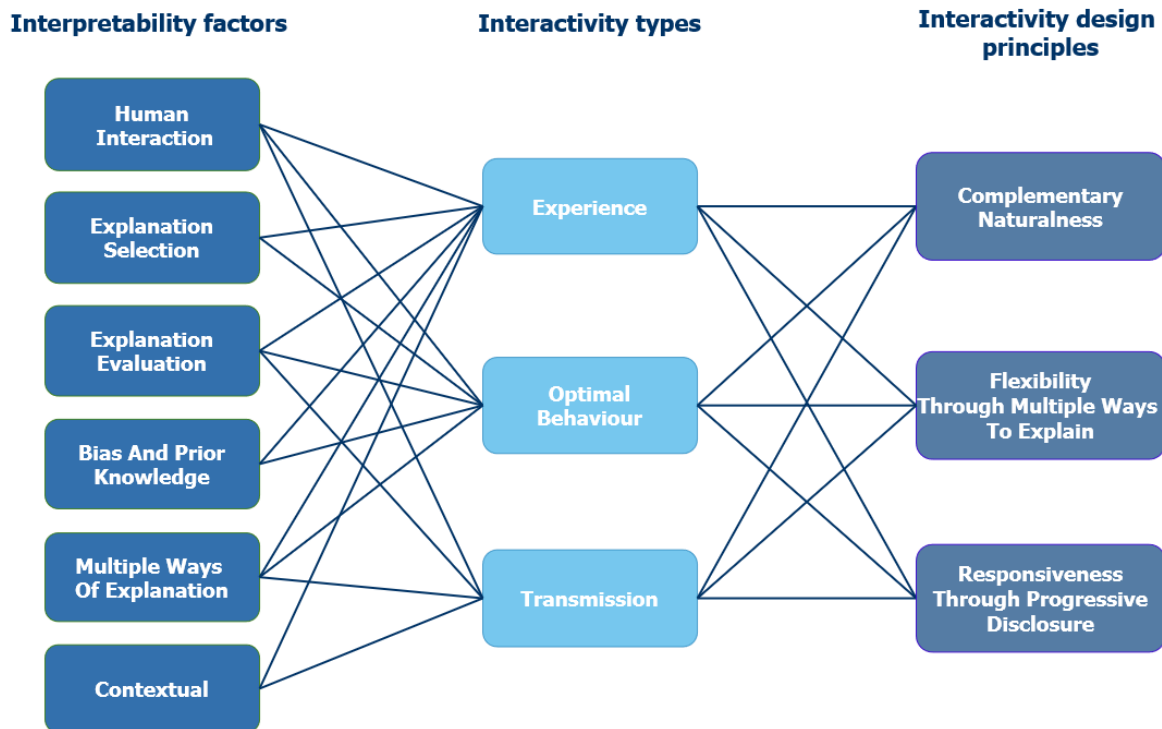


Figure 4.7: Theoretical framework diagram from literature results

In addition to the theoretical framework diagram, Table 4.1 and Table 4.2 further elaborate the connections among the concepts with textual explanations. Table 4.1 outlines how each type of interactivity is associated with specific design principles, providing a clear reference for understanding the practical implications of these principles within the XUI context. This table serves as a crucial tool for visualising the direct relationships between interactivity types and their corresponding design principles, offering a structured overview that guides the design process.

Similarly, Table 4.2 presents a comprehensive mapping of how interpretability factors are linked to their respective interactivity types. This table demonstrates the practical application of theoretical concepts, showing how different interpretability factors enhance or influence the selected types of interactivity.

Together with the diagram in Figure 4.7, the tables help clarify the theoretical foundations of the XUI design, demonstrating a clear linkage between theory and practical application.

Table 4.1: Interactivity Types and Linked Design Principles

	Complementary naturalness	Responsiveness through progressive disclosure	Flexibility through multiple ways to explain
Interaction as experience	The experiential aspect of interaction is enhanced by complementary naturalness, which ensures that the user not only experiences the system's functionality but also receives explanations in a manner that feels intuitive and reassuring.	Tailoring interactions based on experience means progressively disclosing information in response to user queries, adapting to their comprehension level and managing their expectations through the flow of interaction.	By offering a variety of explanatory methods, user can experience the AI system through different lenses, whether through interactive models, narrative explanations, or visual graphics, each enriching the user's experience and understanding.
Interaction as optimal behaviour	When focusing on optimal behavior, combining visual and verbal explanations allows users to navigate the system more efficiently, ensuring that the interaction is not only optimal in function but also in understanding.	This principle plays a significant role in fostering optimal behavior. As users interact with the system, they are provided with explanations at a pace and depth that they can handle, encouraging efficient and informed engagement.	Optimal behavior is supported by providing users the freedom to explore explanations in a manner best suited to them, be it through data manipulation, scenario testing, or simply consuming information in their preferred format.
Interaction as transmission	In information transmission interactions, the XUI acts primarily as a conduit for explanation. However, when paired with the complementary naturalness principle, the interface can enhance this basic exchange by supplementing visual outputs with accessible, natural language rationales, ensuring that explanations are not only transmitted but also understood.	With information transmission, the initial data provided might be quite straightforward. Integrating the principle of progressive disclosure can enable a deeper exploration upon user request, allowing for subsequent layers of detail to be revealed in an easy-to-follow, interactive manner.	Even in the simplest form of interaction, offering various methods of explanation (visual aids, detailed breakdowns, summaries) can accommodate different user needs. This aligns with the flexibility principle by allowing users to choose how they wish to receive and process information from the AI.

Table 4.2: Interpretability Factors and Their Corresponding Interactivity Types

	Interaction as experience	Interaction as optimal behaviour	Interaction as (inter)action transmission
Human interaction	In this interaction type, the focus on clear and direct communication aligns with the need for AI systems to provide explanations that are intuitively understandable and mirror human-like explanation patterns. This ensures that the explanations resonate with natural human reasoning, making them more relatable and easier for users to evaluate.	Emphasising human-like interactions and continuous feedback, this type mirrors the natural social dynamics of human interaction. By providing a responsive and adaptive explanation experience, it meets the human expectation for engaging and relatable communication, enhancing the overall interpretability of the system.	This interaction type involves tailoring the system to user behaviors and expectations, closely aligning with the human desire for explanations that fit within their cognitive and social contexts. By optimising how information is delivered to match user preferences and understanding, it ensures that explanations are not only accessible but also resonate on a personal level with users.
Explanation selection	Interaction as experience involves selecting the most pertinent information to present based on user preferences. This selection is guided by understanding what the user finds abnormal, their intentions, and the necessary and sufficient conditions for an outcome, thereby enhancing the relevance and impact of the explanations provided.	When examining limitations in user interaction, factors such as contrastive explanations, abnormal events, and the intentions behind actions can be highlighted to guide users towards understanding the AI's decision-making process, thus optimising their behavior in response to the explanations provided.	
Explanation evaluation	Interaction as experience focuses on providing clear and coherent explanations, which is essential for effective explanation evaluation. By presenting information that aligns well with users' existing beliefs and simplifies complex data, it supports the evaluation criteria of coherence and simplicity.	In this interaction type, managing user expectations about the accuracy and limitations of AI systems influences how explanations are evaluated. By maintaining transparency and adapting to user feedback, it fosters a deeper understanding and trust in the system's outputs.	This interaction emphasises tailoring the system's behavior to enhance user interactions, which directly impacts how explanations are assessed. By optimising how information is presented and making it relevant to the user's context, it ensures that explanations are both practical and applicable, aiding in their evaluation.
Bias and prior knowledge	Interaction as experience must account for user biases and prior knowledge. Explanations need to be crafted considering these factors to ensure they are perceived as relevant and trustworthy. Biases might shape the experience by influencing which explanations are more likely to be accepted by the user.	Interaction as optimal behavior must navigate user biases and prior knowledge to direct their attention to the most relevant and accurate explanations, aiding in their understanding of AI decisions and potentially reshaping their biases towards a more accurate interpretation of the AI's capabilities.	

<p>Multiple ways of explanation</p>	<p>This interaction type underscores the utility of diverse presentation formats. By offering explanations through visual, textual, or auditory means, it caters to various learning styles and comprehension levels, enhancing the transparency and accessibility of information.</p>	<p>Emphasising varied explanatory methods enhances user experience by adapting to individual learning preferences and ensuring interactions feel intuitive and engaging. This method makes the interaction more reliable and effective by mirroring natural human communication styles..</p>	<p>Incorporating multiple explanation modalities can guide users toward more effective interactions with the system, allowing them to explore different aspects of AI operations. This approach helps users navigate the system more efficiently, making informed decisions based on a fuller understanding of AI outputs.</p>
<p>Contextual</p>	<p>Since the need for explanations varies depending on the user context, interaction as experience should adapt to the level of detail required by different user groups. It should also consider the specific domain context to ensure explanations are relevant and applicable to the end-users' real-world problems.</p>	<p>Optimal behavior interaction is sensitive to the context in which explanations are delivered. By adjusting explanations based on the individual user's context, including their domain knowledge and cognitive styles, interaction becomes more personalised and effective, ultimately leading to enhanced interpretability and a more intuitive user experience.</p>	<p>Information transmission must consider the context in which it is delivered to ensure interpretability. The XUI should adjust the complexity and depth of information based on the user's context, such as their domain expertise or familiarity with the AI system.</p>

4.2.4. Conclusion

In conclusion, this subchapter has effectively answered the sub-research question :

"How can interpretability and interactivity be linked to design principles for an XUI?"

by aligning the theoretical insights from existing literature with the practical requirements of the FOKUS use case. The collaboration with FOKUS defined the scope of the XUI through a detailed exploration of XAI methods and the conceptualisation of design elements that would be both functional and intuitive for medical professionals, as detailed in section 4.2.2.

The process of addressing essential questions—what to explain, how to explain it, and to whom—clarified the scope and limitations of the FOKUS project and the needs of the target audience. This clarification enabled the selection of appropriate interactive design principles: complementary naturalness, responsiveness through progressive disclosure, and flexibility through multiple ways to explain. The deliberate exclusion of the sensitivity to context and mind principle from the initial prototype reflects a strategic choice based on the current capabilities of the system.

The theoretical framework, depicted in Figure 4.7 and further detailed in Tables 4.1 and 4.2, effectively bridges the theoretical concepts of interpretability and interactivity with practical design principles. This framework outlines how interpretability factors are interconnected with various types of interactivity and grounded in design principles tailored to the specifics of the use case.

Conclusively, this chapter has laid a robust theoretical foundation for designing an XUI that integrates essential aspects of interpretability, interactivity, and design principles. By methodically connecting these components, the theoretical framework essentially provides a comprehensive roadmap for the development of the XUI that not only meets the technical requirements of XAI but also resonate with the cognitive and practical realities of end-users in the medical field.

4.3. XUI Design

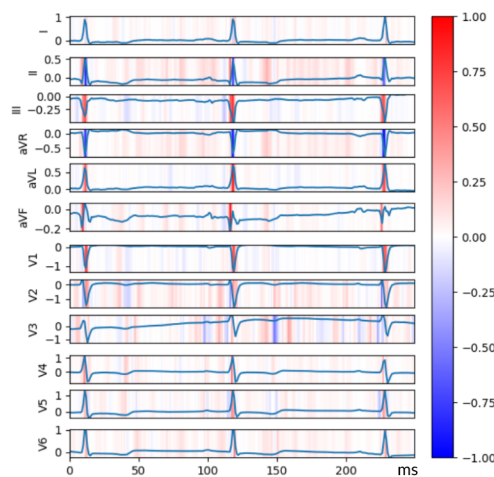
This section will describe how each design principle will be implemented in XUI, the process of designing and implementing each XUI element and showcase the final XUI design.

4.3.1. XUI elements

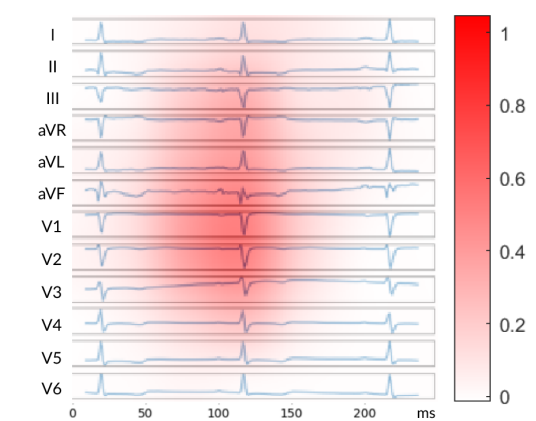
XAI methods

The selection of three XAI methods for display in the XUI was carefully considered to balance breadth and depth of analysis without overwhelming the user. A choice between only two methods might have seemed limited to a mere comparison, whereas four could dilute focus and complicate interpretation. Thus, three methods—SHAP, LIME, and Grad-CAM with varying colour scales—were chosen to provide diverse perspectives on the same diagnostic data, enhancing the robustness of the analysis. The outputs of these methods were previously described section 4.1.2.

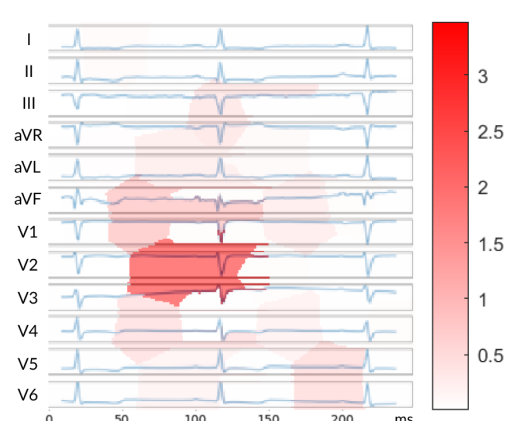
To maintain consistency and clarity in the presentation, several formatting standards were applied across all three methods. Each method's output was arranged to show all ECG leads marked on the left side of the image, time progression at the bottom, and a colour scale with values on the right. Only the SHAP output included an additional numerical scale for the leads on the left side due to the XAI method. The colour coding was standardised as well: red to indicate features positively impacting the AI's output, and blue for those with a negative impact. These uniform adjustments, detailed in Figure C.1, ensure that the differences between the methods are clear, focusing attention on the interpretative nuances each method brings to the diagnostic process.



(a) SHAP



(b) Grad-CAM



(c) LIME

Figure 4.8: Chosen XAI methods

Patient selection

To demonstrate how XAI methods can vary across different patients, two distinct patients were selected to be featured in separate but identically laid out XUIs. This approach aims to illustrate the variability in ECG data presentation for different patients within the same XUI framework. During earlier discussions with the FOKUS research team, a shortlist of 13 patients was already prepared from the larger PTB-XL dataset, which includes 18,885 patient records (Wagner et al., 2020). Since the XAI outputs for these 13 patients were already generated, the selection was narrowed to this group to utilise the existing data, rather than starting anew from the full dataset without prepared outputs. The first patient was chosen based on the complexity and activity evident in their visual outputs, while the second patient was selected for their outputs' contrast to those of the first patient and having the opposite sex. Both selected patients were identified as having an MI according to the dataset annotations, and similarly, the FOKUS AI model also marked them as having an MI. For a comprehensive comparison, Appendix C features a side-by-side view of the XAI outputs for the two selected patients across all three XAI methods.

Explanation elements

This section will explain the different type of explanation elements added to the XUI. The explanation elements which are derived from the (X)AI model were defined during the analysis of the scope of the XUI in collaboration with the FOKUS researchers, detailed in section 4.2.2. Further, during the design phase, the explanations provided for the XAI methods were rigorously evaluated with input from the FOKUS team to confirm their accuracy and appropriateness in describing the model's functionality. This process ensured that all explanations added to the XUI are both factually correct and effectively clarify the XAI models.

Patient information

This element was introduced to provide the background information of the patient, offering context but not directly explaining the XAI method. This additional information, which includes the patient's ID, age, sex, height (in cm), weight (in kg), and the date/time (formatted as DD/MM HH:MM without the year), is crucial for understanding the ECG data. It could influence the user's interpretation of the data displayed. The date excludes the year to more realistically simulate real-time scenarios, given that the patient data in the dataset was collected between 1989 and 1996 (Wagner et al., 2020). The patient information element is displayed in Figure 4.9.

Patient information	
Patient ID	7591
Age	50
Sex	Female
Height (cm)	175
Weight (kg)	83
Date/Time	26/04 09:10

Figure 4.9: Explanation element: Patient information

XAI method

This element was introduced to provide an textual explanation for each of the the XAI methods used. Initially, draft explanations were generated with the assistance of ChatGPT3.5, using the following prompt adapted for each specific method:

Can you write a short paragraph about the XAI method ... if you were explaining it to someone who has no expertise in this area?

The preliminary generated explanations were subsequently reviewed by a diverse group of individuals, varying in their familiarity with (X)AI, to gather feedback and refine the content. Each explanation was intentionally refined to progressively simplify the jargon used, starting with SHAP, followed by Grad-CAM, and finally LIME, while incorporating distinct information relevant to each method. These refined

explanations for each XAI method are presented in Figure 4.10, providing accessible insights into how each method works and how the visual output was generated.

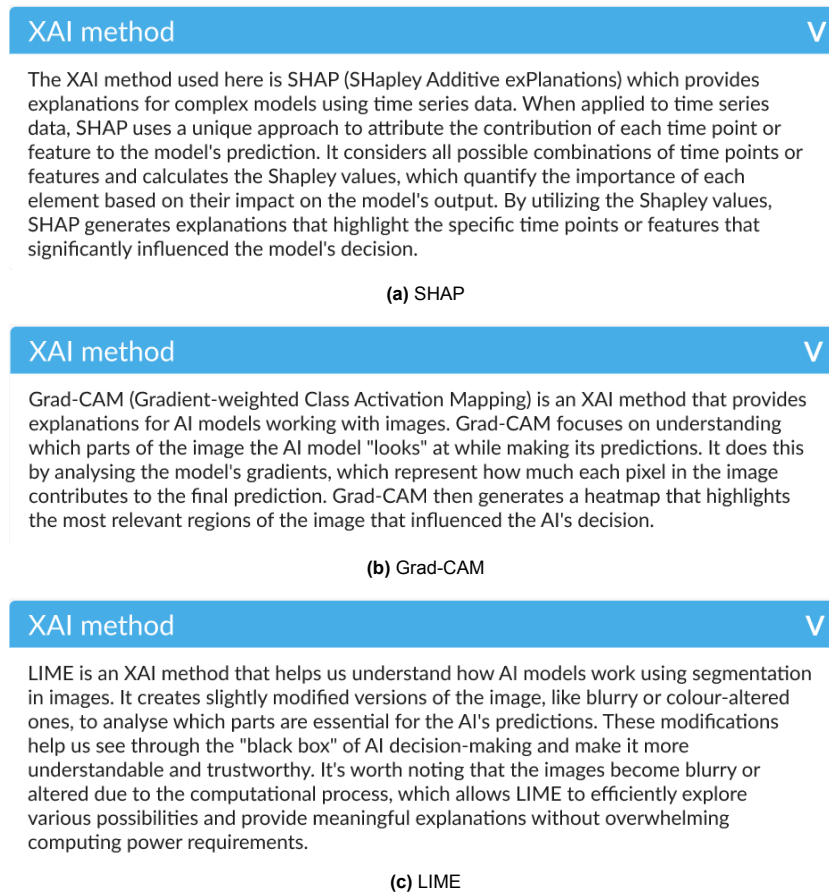


Figure 4.10: Explanation element: XAI method

Prediction score

The element was introduced to provide the user with a numeric prediction score from the AI used to determine the likelihood of an MI. This score is an additional output from the FOKUS (X)AI models and offers a quantitative assessment of the AI's confidence in its diagnosis. Literature suggests that reliance solely on probabilities can be unsatisfactory (Josephson & Josephson, 1996; McClure, 2002; Miller, 2019); therefore, alongside the numeric score, a textual explanation is provided, clarifying the significance of the score in plain language. To enhance readability and emphasis, the textual description of the prediction score is highlighted in bold. This layout contributes to a more intuitive user experience by balancing the visibility of crucial information within the interface. The prediction score element is displayed in Figure 4.11.

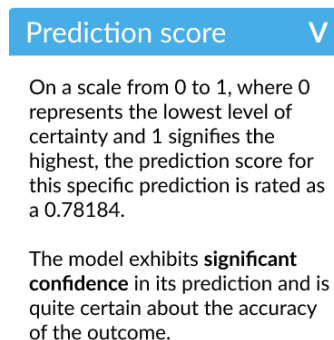
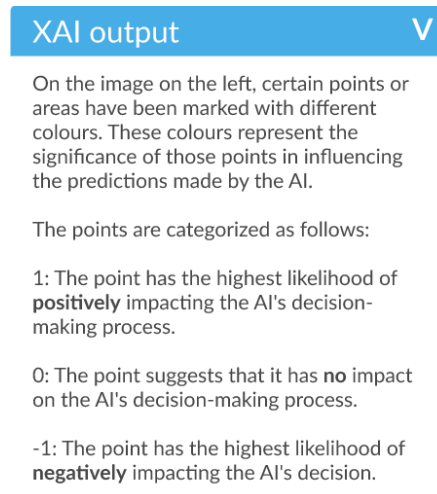


Figure 4.11: Explanation element: Prediction score

XAI output

The element was introduced to explain the colour markings in the output of the XAI method. The explanation is presented through a combination of numerical data and descriptive text, explaining the significance of the colour(s) and the(ir) impact on the XAI output. This textual explanation complements the visual information, providing users with an additional perspective. Key terms indicating the extent of the colors' impact are highlighted in bold, improving both readability and emphasis. The XAI output element is displayed in Figure 4.12.



XAI output ▼

On the image on the left, certain points or areas have been marked with different colours. These colours represent the significance of those points in influencing the predictions made by the AI.

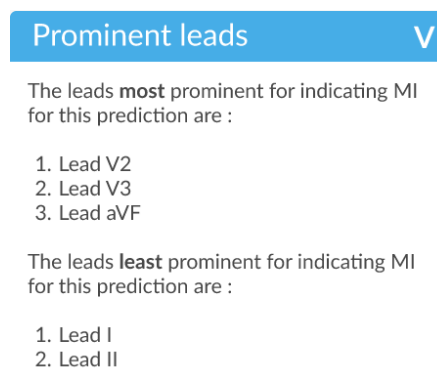
The points are categorized as follows:

- 1: The point has the highest likelihood of **positively** impacting the AI's decision-making process.
- 0: The point suggests that it has **no** impact on the AI's decision-making process.
- 1: The point has the highest likelihood of **negatively** impacting the AI's decision.

Figure 4.12: Explanation element: XAI output

Prominent leads

The element was introduced to direct users' attention to specific areas in the image that the XAI identifies as most or least influential in the AI prediction. This guidance could be useful in the examination of the output, especially for aspects that may not be immediately apparent to the unaided eye. By presenting this information alongside the visual output, users are encouraged to consider parts of the data they might otherwise overlook, potentially altering their interpretation of the XAI's analysis. The explanation is delivered in textual form, with key terms indicating the impact on the prediction highlighted in bold to enhance readability and emphasise significant points. The prominent leads element is displayed in Figure 4.13.



Prominent leads ▼

The leads **most** prominent for indicating MI for this prediction are :

1. Lead V2
2. Lead V3
3. Lead aVF

The leads **least** prominent for indicating MI for this prediction are :

1. Lead I
2. Lead II

Figure 4.13: Explanation element: Prominent leads

Design integration of the XUI

The final design of the XUI layout emerged from an iterative process that drew upon conventional principles of user interface design, the practical aesthetics of medical interfaces, and a dash of creativity.

At the core of the XUI lies the XAI output, occupying the central position to anchor the user's attention where it is most needed. Surrounding this focal point, complementary elements are positioned at the sides to form a coherent whole. The interface adopts a colour scheme rooted in the clinical neutrality of blues and greys commonly associated with healthcare environments, ensuring a familiar and non-intrusive user experience. Strategic use of contrasting colours differentiates interactive sections, guiding the user through the XUI with intuitive visual cues.

The XUI's layout evolved from a dynamic process of design iteration, with the menu positioned at the top to maximise horizontal space and facilitate user navigation. Selecting among the explanation options is streamlined, with the active choice underlined to signify current engagement. Additionally, an alert for the detection of MI is prominently placed at the right, coloured in red, to immediately communicate critical information.

Organisation of content within the XUI was approached with readability as a priority. With the broader explanation of the XAI method granted the most horizontal space due to its descriptive nature. Vertically oriented spaces flanking the central display are optimally used for more bullet-style information. Aligning with standard medical UI practices, patient information is positioned on the top left, while an explanation of the XAI output logically sits to the top right, offering a direct visual connection to the corresponding area of the central output which is situated at the right side of the output.

The lower sections of the interface—left for prediction scores and right for prominent lead identifications—are filled without specific functional rationale, instead serving to complete the explanation offering of the XUI.

The XUI is fully displayed in Figure 4.14 for patient 1 and explanation 1. Appendix x features the remaining explanations for patient 1 and all the explanations for patient 2.

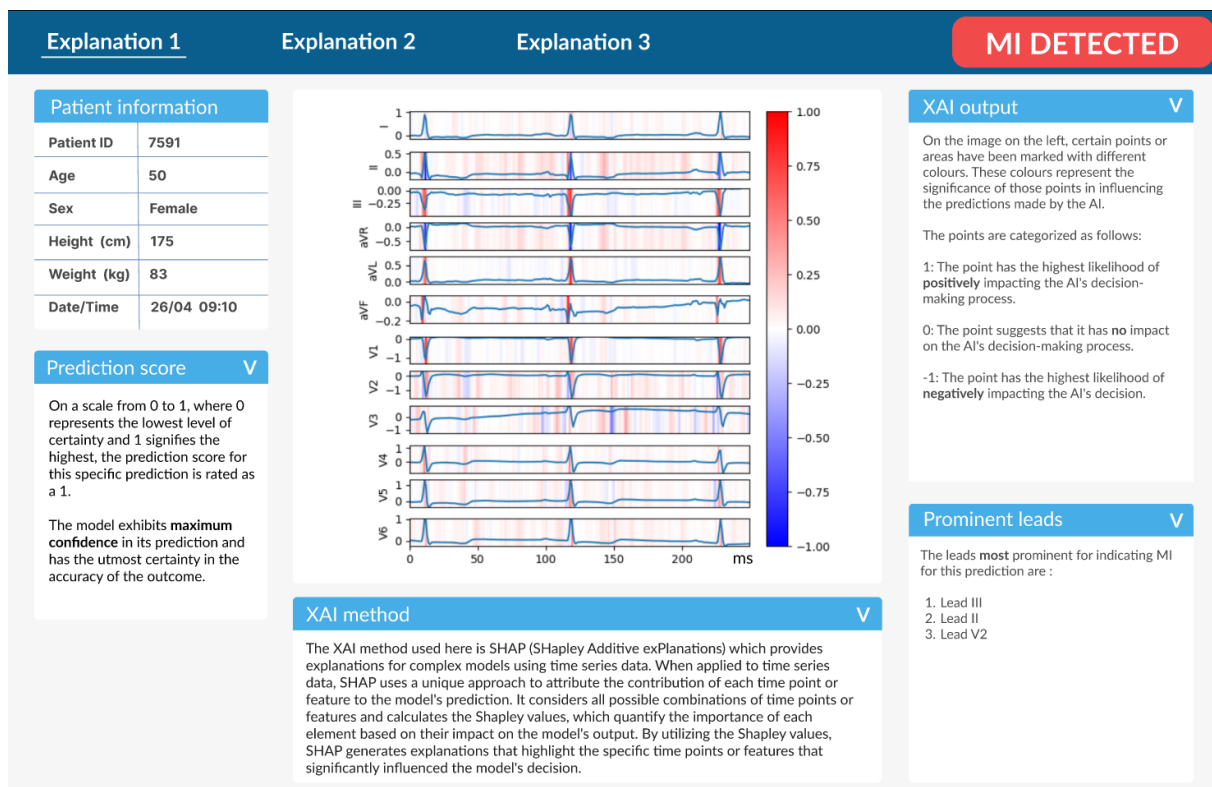


Figure 4.14: XUI displaying explanation 1 for patient 1

Disclosure of the elements

The disclosure of the elements was determined by the importance of the explanations. The XAI output and patient information are presented as always-visible core elements, reflecting their integral role in the interface's functionality. These key elements are not collapsible, thereby ensuring continuous access to vital information central to the XUI's purpose by remaining disclosed.

Contrastingly, the remaining elements of prediction score, XAI method, XAI output and prominent leads are initially hidden, deeming them non-essential for immediate attention but available upon user interaction. Interactive bars allow users to reveal or conceal these elements at will, with intuitive visual indicators at the top right of each bar signifying their current state. This design facilitates an environment of controlled exploration, allowing users to manage the disclosure of information based on their individual needs.

When users access the XUI, they will first encounter Explanation 1, with the XAI analysis and patient information already revealed. If users navigate to other explanations for the first time, they will find these elements remain consistently disclosed. Additionally, the open or closed status of other elements is preserved, reflecting the user's last interaction when toggling between different explanations.

Furthermore, the user interface is enhanced with cursor-change feedback when interacting with the collapsible elements, signifying the possibility of interaction and inviting users to engage with the content.

Figure 4.15 depicts the open and closed state of the element prediction score. Figure 4.16 depicts the initial presentation of the XUI when upon user engagement.

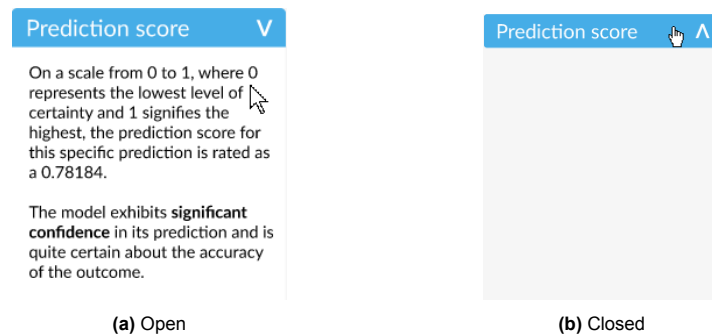


Figure 4.15: Disclosure of element prediction score in open and closed state

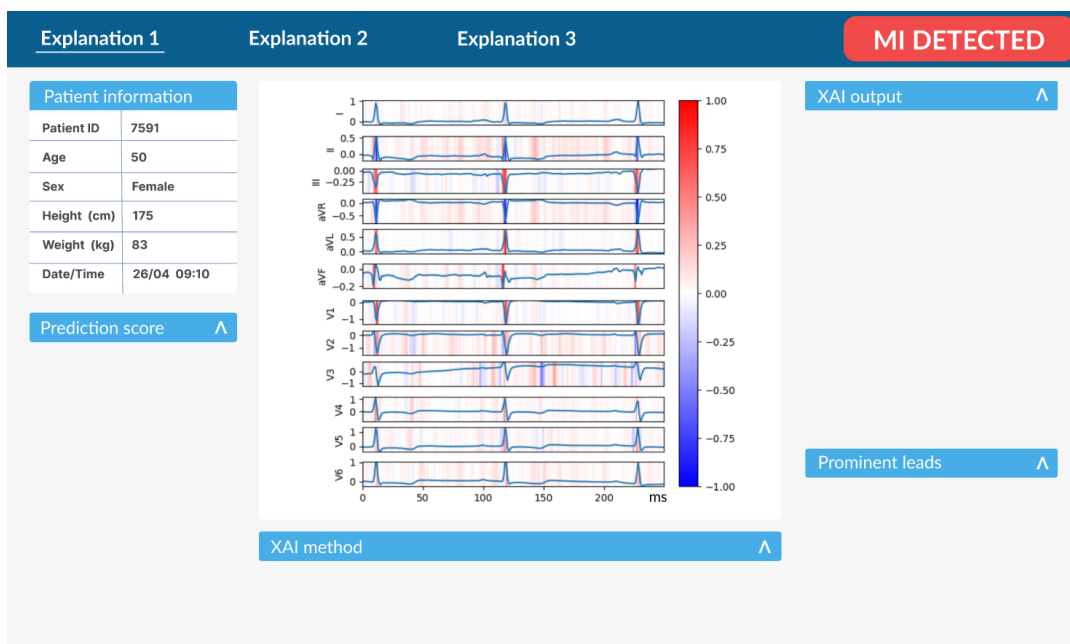


Figure 4.16: Initial presentation of XUI displaying explanation 1 for patient 1

4.3.2. Design principles applied

In the design of XUI, the adoption of design principles is applied to align with the user's needs for interpretable explanations of the XAI methods and to assure that the design was grounded in theory. The implementation of these design principles within the XUI is as follows:

Complementary naturalness

- Textual explanations accompany visual representations, providing a verbal narrative that explains the underlying XAI methods in a way that complements the visual data.
- A variety of textual descriptions are provided for different aspects of the XAI output, offering users multiple angles from which to interpret the output, including textual elaborations on the XAI method, the predictive score, and the prominent leads.

Flexibility through multiple ways to explain

- To offer a multifaceted perspective on diagnosis, the XUI implements three distinct XAI methods for analysing the same patient data, providing a flexible viewpoint.
- The XUI combines visual and textual explanations, allowing users to see the XAI outputs and read about them in detail, aiding in the comprehension of the visual information.
- The explanations include both numerical data and their textual interpretations. This is achieved by offering a textual explanation alongside a numeric score clarifying the significance of the score in plain language.

Responsiveness through progressive disclosure

- Elements within the XUI can be expanded or collapsed by the user, providing control over the flow and quantity of information consumed, which is particularly useful in managing cognitive load.
- Some elements are set to a default closed position to streamline the user's focus and prevent information overload, while key information remains constantly visible to guide the user's attention to essential details.

4.3.3. Conclusion

This sub-chapter has addressed the sub-research question:

"How can interactive design principles be applied to design an XUI for end-users?"

By detailing the implementation of design principles within the XUI developed for the FOKUS use case. This process illustrated how theoretical principles could be practically applied to enhance the interpretability of XAI outputs for end-users.

The selection and integration of three XAI methods—SHAP, LIME, and Grad-CAM—within the XUI was strategically implemented to offer a well-rounded and comprehensive analysis without overwhelming the users. This integration, coupled with the additional explanation elements such as patient information, prediction score, XAI method, XAI output, and prominent leads, exemplifies the principle of **Flexibility through multiple ways to explain**. This approach ensures that users receive varied perspectives on the AI's diagnosis, to enhance the understanding through multiple interpretive angles.

The integration of textual explanations alongside visual data employs the principle of **Complementary naturalness**. This approach ensures that users receive a coherent narrative that explains complex XAI outputs in an accessible manner, enhancing the naturalness and intuitiveness of the information presented.

Moreover, the principle of **Responsiveness through progressive disclosure** was adeptly implemented, allowing users to dynamically interact with various XUI elements. This feature enables users to reveal or conceal specific explanation elements at will, tailoring the interface to their cognitive preferences and needs. By setting some elements to be disclosed by default and others not, the design effectively guides users towards the most crucial data, preventing information overload while maintaining user engagement and interaction.

The application of these design principles not only catered to the functional requirements of the XUI but also addressed the cognitive processes of the end-users, ensuring that the system is user-friendly and effectively supports decision-making processes in the context of the use case.

In conclusion, the chapter demonstrates a structured approach to applying interactive design principles in the design of an XUI, ensuring that the system is not only functional but also intuitive and responsive to the needs of its users.

5

Validation XUI

5.1. Data Collection and Analysis method

As discussed in earlier chapters, the research methodology utilised semi-structured interviews for their flexibility and investigative potential. To effectively validate the design of the XUI, 8 specialists who are active in the medical field were interviewed, with the visualisation of the thematic analysis depicted in Figure 5.1. Thematic analysis was utilised to analyse the qualitative data, a method renowned for its efficacy in identifying, analysing, and reporting patterns within data (Braun & Clarke, 2006). Subsequent sections will discuss the overview of the interview process.

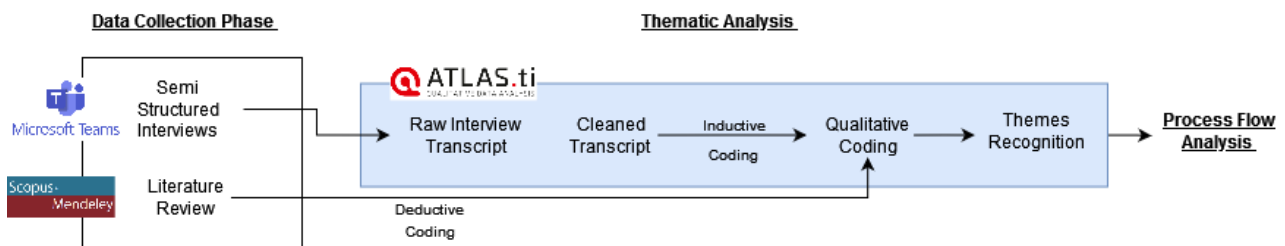


Figure 5.1: Qualitative Data Analysis Process

5.2. Sampling Process

Before starting the qualitative research, establishing a strategic sampling plan is crucial for ensuring that the most relevant data is collected within the constraints of available resources. Inspired by Robinson (2014), this study adopts a purposive sampling approach, designed to selectively include participants who offer insights and firsthand experiences from the medical field. The goal is to select individuals whose contributions will enrich the depth and variety of the data collected. Table 1 outlines the sampling strategy adopted for this research.

Phases in Sampling	Description
Identification of Target Participant Pool	Individuals with background in cardiology, either directly related (MD) or associate doing PhD research
Selection of Sample Size	8 participants
Strategy for Sampling	Convenience Sampling
Recruitment of Sample Participants	Snowball Sampling

Table 5.1: Sampling Process

While narrowing down the problem statement and research questions, the primary inclusion criteria were set as individuals with background knowledge in cardiology. This selection encompassed both direct experts, such as cardiologists holding MD degrees with proficiency in ECG analysis, and scholars in the field of cardiology who, despite lacking an MD, have substantial affinity or experience with ECGs. This deliberate inclusion of both profiles aimed to capture insights from both perspectives: the practicing cardiologist, who is the primary end-user, and the non-MD researcher, offering their perspective of the field. With these selection criteria in place, 8 interview participants were recruited using a snowball sampling technique. The participants were primarily reached out to through personal connections within professional networks.

5.3. Interview Process

Participant	MD background	ECG knowledge	MI determination	Specialisation
P1	MD	Advanced-Expert	Yes	Sports Cardiology
P2	Non-MD	Novice	Very Limited	PhD in Experimental Cardiology
P3	MD	Advanced-Expert	Yes	Genetic cardiomyopathy (Cardiology)
P4	MD	Expert	Yes	Cardiology
P5	MD	Advanced-Expert	Yes	Sports Cardiology
P6	Non-MD	Novice	Very Limited	PhD in Experimental Cardiology
P7	MD	Average-Advanced	Yes	General Practitioner
P8	MD	Advanced-Expert	Yes	Cardiology

Table 5.2: Interview Participants Overview

Before the commencement of the interviews, a detailed interview protocol was devised, to align with the features of the XUI and the overarching themes of the research topic. This protocol was structured to maintain consistency across all interviews while allowing for a thorough exploration of the relevant research question. At the beginning of each interview session, participants were briefed on the study's objectives and the expected outcomes of their participation. This introduction aimed to set clear expectations and encourage candid and informative responses.

To gather comprehensive background information, initial questions focused on the participants' roles and experiences in cardiology, particularly in the ability to diagnosing MI. Following this, the interview progressed according to the predefined protocol. It included specific questions aimed at evaluating the effectiveness implemented design principles and various elements of the XUI on interpretability, as described in section 4.3. The protocol encompassed both targeted inquiries, such as those regarding the use of textual explanations or the inclusion of multiple explanatory approaches, and more open-ended questions to facilitate broader discussions. The full interview protocol is outlined in Appendix A.

The interviews were conducted virtually, using secure online platforms to ensure accessibility and convenience for all participants. Each session varied in length, with an average duration of approximately 45 minutes, ensuring thorough exploration of the topic without imposing excessively on the participants' time. In line with ethical standards and data protection regulations, participants were assured of their anonymity and the confidentiality of their responses. They were also informed of their right to withdraw from the study at any point.

To ensure methodological integrity, this study focused on the reliability and completeness of data by proactively identifying and addressing potential gaps in the research process. A thorough participant selection procedure was adopted, where individuals were selected based on their background in cardiology, guaranteeing a comprehensive grasp of the subject, with all participants possessing relevant experience in this field. At each interview's start, a review of job roles confirmed their relevance to the study's key areas. Furthermore, the research adhered to strict ethical standards, including GDPR compliance and the ethical guidelines of TU Delft, boosting the trustworthiness and relevance of the findings. This methodical selection and interview process reflect the study's dedication to a nuanced comprehension of XAI interpretation in healthcare, with a special emphasis on cardiology.

5.4. Interview data analysis

After collecting the interview data, including recordings and transcripts, the information was organised and analysed using Atlas Ti and the thematic analysis framework by Braun and Clarke (2006).

This approach took a systematic route to sort through the data, identifying themes in six phases for a comprehensive analysis. By applying the methodologies of both grounded theory and thematic analysis, a multifaceted perspective on the process flow can be visualised.

Process Stages	Description
Acquainting with Data	Engaging in multiple reviews of the transcribed data to identify initial coding opportunities.
Initial Code Generation	Utilising identified ideas to compile a preliminary collection of codes.
Theme Investigation	Developing a thematic framework from the established codes to understand the overarching themes.
Theme Assessment	Conducting a thorough examination and refinement of the themes to capture their core significance.
Theme Definition and Labelling	Conducting deeper analysis and evaluation of themes for precise definition and naming.
Report Compilation	Crafting a comprehensive summary of the findings from the interviews, linking them directly to the central research question.

Table 5.3: Process stages of a Thematic Analysis (Inspired by Braun and Clarke (2006))

The analysis of the interview data was a combined approach of inductive and deductive coding to thoroughly examine the data. Deductive coding started with a set of pre-defined codes based on the theoretical framework as concluded per the literature review, focusing the analysis on specific segments of the data. This method applied a theoretical lens to the data, ensuring depth and relevance.

Parallel to this, an inductive coding strategy was adopted, where themes and codes were allowed to emerge naturally from the data, without the initial guidance of pre-existing theories. This bottom-up approach facilitated the discovery of new insights and patterns, making it an invaluable strategy for exploring areas of the research that are novel or less well-defined.

This method of analysis enhanced the organization and interpretation of collected data, enabling a clearer identification of patterns. It streamlined data management, facilitating the recognition of connections among codes and the emergence of relevant themes.

Additionally, this approach was instrumental in articulating the research findings, thereby strengthening the overall validity and depth of the study. The implementation of the thematic analysis framework laid a solid foundation for the study's conclusions, ensuring the themes derived accurately represented the qualitative data's richness.

By employing a combination of deductive and inductive coding strategies, 47 codes were identified. Initially, a broader set of codes was established, but through multiple rounds of analysis, these were combined or filtered out for redundancies and relevance, ultimately refining the list to the final 45 codes. A full list can be found in Appendix B.1 Following the thematic analysis phases, themes were derived using both deductive and inductive methods, starting with the design principles as core themes and subsequently identifying additional themes derived from the interview data.

The culmination of this methodical approach resulted in the identification of 8 primary themes (Table 5.4). These themes share overarching codes, which correspond with the theoretical framework laid out from the literature review, ensuring that the findings were not only empirically supported but also theoretically grounded. This alignment underscores the research's adherence to a rigorous analytical process, affirming the validity of its conclusions and enhancing its contribution to the field.

Table 5.4: Thematic Analysis Framework

Code	Frequency	Complementary Naturalness	Responsiveness Through Progressive Disclosure	Flexibility Through Multiple Ways To Explain	Sensitivity to context and mind	Problems with XAI	Environmental setting of the XUI	Diagnosis	XUI design
Bias/prior	12				x				
Bigger image	2	x				x			
Blurry visuals	6	x				x			
Context	49				x	x	x		
Contrasting outputs by the XAI	12			x					
Counterfactual	6			x					
Decision under uncertainty regardless of outcome	5							x	
Deeper explanations	14	x	x		x				
Deeper underlying issues with AI	8					x			
Difficulty because of ECG	4					x			
Difficulty understanding XAI	12					x			
Disagreement in the diagnosis by AI	14					x			
Dislike for explanation 2	4				x				
Dislike for explanation 3	6				x				
ECG data used is confusing	16					x			
ECG knowledge	7				x				
Engineer perspective	2				x			x	
Experience	14				x				
Explanation selection	25				x				
Guiding on where to look	26		x						x
Lacking ECG knowledge for MI diagnosis	2							x	
Layout	3								x
Multiple explanations	27	x	x	x				x	
Multiple ways of explanation can be confusing if the explanations don't match up	7			x					
No diagnosis based on unreliable output	6					x			
Poor visualisation	24	x							
Preference	15			x	x				
Preference for colour instead of numbers	2	x		x					
Preference for explanation 1	9				x				
Preference for explanation 2	3				x				
Preference for explanation 3	4				x				
Progressive disclosure	24		x						
Simplicity	21	x	x						
Standard ECG visualisation	13					x			
Text	31	x	x	x					
The XAI output does not make sense	34					x			
The XAI/AI is focussed on the wrong things	25					x			
Too much information	8	x	x	x					
Underlying trust	3				x				
Understanding of output	13							x	
Unnecessary information	9	x	x	x					
Usage	18						x		
Use of different colours	3	x		x					
Useful for first time use	24						x		
Visualisation	12	x		x					

5.5. Results

This chapter will discuss the results from thematic analysis as described in the previous section. The results will be described per theme.

5.5.1. Themes

An overview of all the themes, their corresponding codes, and descriptions can be found in Appendix B.2.

Table 5.5 illustrates the frequency of codes associated with each theme. Two significant observations can be drawn from this figure. First, the theme "Sensitivity to Context and Mind," although not initially implemented, emerged prominently during the thematic analysis. The codes linked to this theme appeared frequently enough that their total surpassed those associated with the other three implemented design principles: Complementary Naturalness, Responsiveness Through Progressive Disclosure, and Flexibility Through Multiple Ways to Explain.

Secondly, the theme with the highest frequency total is "Problems with XAI," which encompasses the issues or challenges encountered during the interviews regarding the XAI itself. Given its prevalence and relevance to the findings of other themes, this theme will be addressed first in the following section.

Table 5.5: Code frequency within themes

	Totals
Complementary Naturalness	116
Responsiveness Through Progressive Disclosure	121
Flexibility Through Multiple Ways To Explain	101
Sensitivity to context and mind	125
Problems with XAI	135
Environmental setting of the XUI	64
Diagnosis	20
XUI design	28

5.5.2. Problems with (X)AI

In the course of the interviews and thematic analysis, several vital contextual details and problems came to light, significantly impacting the research findings. These issues were exclusively associated with the (X)AI, setting them apart from elements related to the XUI.

Participants often circled back to these problems when attempting to provide feedback on the XUI, despite the difficulties they encountered. The persistent reappearance of these issues throughout the chapter's discussions underlined their importance, making it crucial to address it as the first theme.

To commence the discussion, one quote about the explanations, referring to one of the visual XAI outputs, humorously encapsulates the challenges the participants encountered while interacting with the XUI.

*"This explanation looks really random to me. It looks like settlers of Catan had a party here"-
P3*

Wrongful dataset usage

The first detail pertains to the ECG dataset utilised for the AI. The primary application of the FOKUS AI was to detect a MI in real-time as it occurs and explain this prediction with their XAI. As outlined in Chapter 4, a dataset of ECGs capturing MI, validated by medical professionals, was employed. This comprehensive dataset, accumulating over several years of research, is widely recognised as reliable. However, it became clear during the interviews that this dataset primarily consisted of data on previous MI episodes, whereas the correct medical terminology for the intended use case of detecting MI in real

time, as clarified by the experts, should be "acute MI". The experts further explained that the indicators for "previous MI" and "acute MI" significantly differ.

"The only signs that could indicate a MI are signs of an old infarction. So signs of QA formation, R amplitude progression but that's not something you look for when you look for acute MI in medical practices." - P3

Some where even quick to point out that the incorrect markings could be due to incorrect dataset use.

"I'm assuming there was a very reliable data set as in there was like enough data. I know these things are very data hungry, but what I really am interested in now is what the features are that the algorithm identified because I'm worried that it's picked up on incorrect features which don't align with signs of acute MI." - P4

Continuing this, the method by which the XAI identified areas of interest—aside from emphasizing incorrect features— also contradicted several standard approaches an MD might use to diagnose a MI through ECG analysis. This was pointed out by MD's with varying levels of expertise, including participant P7, who engages with ECGs occasionally.

"When looking at an ECG, it's really a dynamic process as there is not one lead which indicates whether there's an MI or not. There are different leads because it's like a cumulative electric signal, where there is an MI. You have to look at every lead separately to be able to come to a conclusion." - P7

To participant P3, who interacts with ECG's on a daily basis.

"It's interesting to me that the middle heartbeat is the only one that is contributing to the diagnosis of MI, because all the three beats are the same. So I would have expected like markings here, here and here, because an MI is not diagnosed based on one beat, it would be diagnosed on all the beats. You don't just look at one specific beat but you go through them all and compare them." - P3

Building on the issue mentioned by the above quote, the XAI's output appears to contradict itself by highlighting only specific sections and overlooking others, which should also be considered as interconnected indicators of MI. This selective explanation misrepresents the comprehensive approach needed for accurately diagnosing MI, where the entirety of all the heartbeats should be evaluated, not just isolated instances. This aspect caused confusion among the participants.

Standard ECG layout

The second detail that emerged relates closely to the ECG presentation layout, a topic previously mentioned in the Design chapter due to concerns over its legibility. This issue indeed became a significant obstacle for participants as they interacted with the XUI, with many finding the unconventional layout not only unfamiliar but also counterintuitive. This deviation from the norm provoked a range of reactions, predominantly focusing on the discomfort and adjustment challenges it presented.

"It's a bit unusual to look at the ECG like this. Normally we have them a bit larger with different speed settings and we have a 3 by 4 grid. It's bothering me a bit that I have to look at it like this" - P8

The layout's deviation from standard practice notably impacted critical aspects of medical analysis, such as the speed and ease with which a MI diagnosis could be conducted. In high-stake environments of medical diagnosis, such as the use case, where every second counts, the familiarity with the tools and data presentation can significantly influence the outcome.

"The standard 12 leads layout and the 25 millimetre speeds paper speed. Those two things help in getting your conventional way of interpreting your ECG, which makes it quicker and easier, because that's what most of us are used to working with. " - P1

Interestingly, even participants without a medical background, while not explicitly criticising the non-traditional presentation of ECGs, highlighted the importance of maintaining a familiar layout and visualisation for ECG's due to their training and how they are used to reading ECG's.

"...I don't work with such extended ECG's. But in my opinion, the less you change, the better it is for them to understand and also translates because they're trained to look at stuff in a certain way." - P2

Poor visualisation

The final detail revolves around the legibility of the XAI outputs in terms of their visual representation. Beyond deviating from the standard layout, the clarity of the XAI outputs was notably compromised due to the XAI methods used. This blurriness made the underlying ECG traces almost illegible, causing some participants to give up on trying to interpret the XAI as soon as they saw it.

"Okay, but this is useless to me. I see only pixels. So to me the blurry ECG is one reason why I would not trust this solution. Because it's more difficult now to check it and I have to take more effort myself to look at it and to verify. " - P2

For explanations 2 and 3, the XAI actually had an adverse effect on interpretability, as the blurriness of the image impaired the ability to see what the explanation was trying to highlight, specifically the ECG.

"The issue with the 2nd and 3rd explanations is that the XAI generated explanation in the heatmaps are overpowering the actual ECG traces. In all cases the traces should be clearly visible, because that's what we're going to use as reference" - P4

Furthermore, the XAI occasionally highlighted elements unrelated to the ECG traces, adding confusion and reducing trust in the XAI's outputs. This problem extends beyond visual clarity, affecting the accuracy and relevance of the highlighted information, potentially misleading users.

"You could argue that depending on the ECG's that you put in, there are gonna be lines that have nothing to do with the patient. For example this horizontal line here, it's just a separator between the two leads but it is marked as relevant by the XAI. So I'm not sure if this is the best thing to look at per se and it makes me wonder about the validity of the XAI" - P3

5.5.3. Complementary naturalness

Participants appreciated the additional or different insights provided by natural language rationales, which complemented the information conveyed through visual explanations. This combination was helpful in increasing their understanding of why the XAI outputs were presented in a certain way, although they also noted limitations.

"...it helps in understanding how to approach the output of the underlying model. So you get a sense of what is basing its conclusions on so I think that is helpful." - P1

However, the analysis also uncovered a nuanced perspective on the effectiveness of textual explanations. While the descriptive explanations helped clarify the appearance of the XAI outputs, their effectiveness is significantly diminished when paired with low-quality visuals. This limitation hindered the users' ability to thoroughly inspect and understand the image. This challenge was particularly evident with XAI methods 2 and 3, where the poor quality of visuals significantly undermined the utility of textual explanations.

Regarding the necessity of the different textual explanations alongside visuals, responses varied. Some participants deemed the textual information crucial for fully understanding the visual output, while others viewed some of it as redundant, offering no additional insight beyond what the visuals conveyed.

"I would say that all elements that were added were of use and nothing is really, let's say clutter to what's already on the UI on the screen. You need a text to understand the image." - P5

"For me the text boxes are just text versions of the image. Like this one (pointing to the most prominent leads explanation), they explain the same thing and in my opinion the image already showed the most important parts." - P8

Discussion among participants also touched upon the balance between the quantity of text, its optimal amount, and its complexity. Some participants indicated that the extensive length of the text prompted them to skim through, causing them to overlook crucial details they later deemed useful. Conversely, a subset of participants indicated a preference for more detailed textual explanations, linking this preference to their personal interest in AI. Notably, the participants within this subset exhibited a wide range of familiarity with AI, from minimal to extensive, providing no real indication as to how background could contribute to this preference.

"...this could be maybe a little bit less text, maybe just the score. For me it wasn't clear that there was a number here that was changing every time depending on the XAI method. Because it was such a long text I didn't read it until the end and I missed the number" - P4

"...the text explanations seem complicated as well, some are a bit easier to understand, but that's probably due to the method used. But it really depends on the person, I personally would like more information because I'm interested in AI and I work with it, but I don't think some of my other colleagues would want it." - P8

Preferences for the depth and length of textual explanations varied among participants, reflecting a diversity of needs and perspectives. Nonetheless, a consensus emerged regarding the use of this XUI for regular interactions: participants agreed that both the complexity and length of textual content should be reduced or, in some cases, completely eliminated. This suggests an overarching preference for streamlined, concise textual explanations to enhance usability and comprehension in frequent use scenarios.

"In the long term visual images would be enough. If I saw it as much as 100 times per day, then just the images is already enough" - P7

5.5.4. Responsiveness through progressive disclosure

The feedback on progressive disclosure within the XUI demonstrates the diversity in user preferences and needs, with distinct opinions on the initial state of explanation boxes—whether they should be fully open, fully closed, or a mix of both to balance immediate access with information overload avoidance.

Full Disclosure Preference: A group of participants favored having all explanation boxes open from the start, valuing direct access to comprehensive information without additional interaction required to reveal content.

Selective Disclosure Preference: Another segment preferred starting with all explanation boxes closed, aiming to mitigate feeling overwhelmed by information. This preference for a cleaner, more streamlined initial view underscores a desire for deeper information to be available but not immediately visible.

Hybrid Approach: There was also a preference for a hybrid setup, where certain explanations are immediately visible, while others remain hidden until explicitly accessed. This approach aims to strike a balance, offering some information upfront while keeping additional details behind user-initiated actions.

"Having the information displayed in different bits is very useful, especially the explanation like I said. It is something you will read the first time you see the user interface, but it's a lot and can be overwhelming. Not having it open immediately is definitely a better approach."
- P5

"..I think it would be more helpful if it's 1 screen and everything was already popped up, that would be easier for me to navigate so I can understand and process all the information better" - P7

Despite these varying preferences, a key consensus among participants was the critical importance of having the flexibility to control the opening and closing of explanation boxes on demand. This feature was seen as especially valuable for users anticipating regular interaction with the XUI. Participants noted that while they might not need in-depth explanations on every interaction, having the option to access this information easily when necessary—without it being permanently on display and potentially contributing to visual clutter—was highly appreciated.

"I like the ability to check out the different explanations, it is nice that you can click on it just to check, and if you've read once it's kind of self-explanatory but good to have a refresher."
- P6

5.5.5. Flexibility through multiple ways to explain

The integration of visual, textual, and interactive explanations within the XUI enriched the interpretative experience by offering a broad spectrum of insights. Participants shared varied perspectives on how this diverse array of explanations enhanced their understanding.

One notable reflection was on how certain explanations illuminated aspects of an image that might have otherwise been overlooked or misinterpreted, guiding users to reconsider their initial perceptions. This ability to shift viewpoints was particularly appreciated in instances where the MI diagnosis was in question, underscoring the value of having access to multiple interpretative angles.

"I think the different explanation boxes are also very helpful. For example. the prominent leads gives you a text to which lead is most interested in, which might be different from if I only had to look at the image" - P1

Furthermore, the array of available explanations allowed users the opportunity to select the interpretation that best resonated with their own reasoning. This selection process not only fostered a deeper connection with the material but also enhanced users' confidence in their understanding, as they were able to choose explanations that aligned most closely with their mental model.

"Yes, it's always nice to have different perspectives. I would probably look at all the explanations and see if I would agree with the different outputs. If there is one I agree with more, I would probably trust that one better as well and I would only look at the one and discard the rest. I probably wouldn't even look at them." - P7

"Having more options kind of gives you the room to choose which one you like the most which was helpful. That's that's just how my thinking works best and that would have aided in interpreting these explanations I think." - P5

Despite these advantages, the sheer volume of explanations available could also overwhelm users, detracting from the user experience by complicating the decision-making process rather than facilitating it. Confusion emerged as a significant issue when explanations did not align, either due to differing interpretations of the same ECG data presented by the different XAI methods or a discrepancy between the user's perspective and the explanation provided.

"So if three marked the same point, I would understand it and trust it. If all three differ, then it would confuse me and I wouldn't trust it. At that point I would base it on my own judgement, and I would ignore the the marks." - P6

"I would say 4 XAIs is definitely too overwhelming. Maybe maybe 1 less is also also a good possibility. So then if you feels a bit more like comparing 2 of them." - P5

While multiple explanations can enrich understanding, there exists a delicate balance between offering sufficient interpretative diversity and maintaining clarity and coherence in the XUI.

5.5.6. Environmental setting of the XUI

As highlighted in earlier sections, both the environment and the intended use of the XUI significantly influence its design. All participants, aware of the XUI's purpose, offered adaptable feedback, especially when considering its future or more frequent usage.

When envisioning scenarios involving regular interaction with the XUI, participants' perceptions and feedback diverged from their initial impressions. This distinction emphasised the need for different UI approaches tailored to specific contexts: for first-time interactions or training, versus routine use in clinical decision-making.

For initial interactions or training, participants appreciated the XUI's current design. The absence of time constraints allowed them to absorb the provided information fully, finding the blend of textual and visual explanations particularly helpful for a comprehensive understanding of the XAI. Despite the challenges identified in the section **5.5.2 Problems with XAI** — such as discrepancies in XAI outputs or the focus on ECG areas typically not considered by medical professionals — participants saw potential benefits in these differences for research or training. They suggested that exploring these variations could enhance research on MI detection or familiarise users with various XAI methodologies.

"...and as I read the ECG, I would look at the the T Wave and ST segments first but as you can see here in this explanation, that's not the way AI bases its decision on so that's very interesting and I think it can be useful also to train people" - P5

However, perspectives shifted when considering the XUI's application in a routine clinical setting. In this context, participants anticipated becoming more acquainted with the XUI and, consequently, less reliant on extensive textual explanations. They expressed a preference for a greater emphasis on visual cues accompanied by concise textual guidance to highlight crucial information quickly, especially under the time constraints of a clinical environment. The need for efficiency and directness became paramount, with any feature that could detract from swift diagnosis — such as lengthy explanations or an overabundance of XAI interpretations — deemed unsuitable. This preference underscores the importance of a streamlined, focused approach in high-pressure settings.

"...this would be more useful in training. The first time you use this, you're greeted with these helpful explanations, walking you through what's happening. It's great – it tells you what's good, what's not, all in a straightforward way. But after a while, I'd expect to get the hang of the XAI, to trust it enough that I just need to quickly check in, validate its findings, and move on. At that point, all those extra explanations start to feel like unnecessary clutter, just distractions really." - P1

5.5.7. Diagnosis

During the interviews, participants were asked if they understood the XAI outputs and if they could make an MI diagnosis with the information provided.

The majority stated they understood the XAI's output to an extent, aided by the detailed explanations. They recognised the XAI's markings and how these aligned with the model's rationale. However, their agreement with the MI diagnosis was reserved, primarily due to the use of an incorrect ECG dataset for the task at hand.

When inquired whether they could make a diagnosis with the XAI's information, all but two participants (those without a medical background) affirmed they could. Yet, this was more a reflection of their own expertise than reliance on the XAI, which was criticised for highlighting irrelevant features.

The use of an incorrect dataset also introduced frustration and confusion among the participants. They expressed concern that in a real-time clinical scenario, reliance on inaccurate XAI inputs could be counterproductive, marking time as wasted on misleading information. However, they viewed false positives differently, understanding the critical nature of healthcare decisions and preferring an overly cautious approach.

"Yes, I would have enough information to make a diagnosis and come to the conclusion that it is not (acute) MI. But if this was the case, I would be annoyed by the AI because it was stealing my time with false information and by giving me visual clutter. I would also probably start to doubt myself. Like, wait a second, am I wrong here?" - P4

Additionally, using inaccurate data led some participants to question their own clinical decisions. This wasn't due to the model's tendency to generate false positives, but rather because they would not have personally classified the cases as MI. This situation underscores how incorrect data can deeply affect professionals' confidence and their faith in XAI systems.

5.5.8. Sensitivity to context and mind

The analysis brought to light a fourth design principle that was left out during the design phase, emerging organically despite not being formally implemented. Reflections on results from previous sections underscored that, although participants shared a common professional background, they displayed distinct individual preferences and interpretations. Although there was a general agreement on the foundational design elements, such as the benefit of integrating textual and visual explanations, the nuances of individual preferences were far more detailed and varied.

To begin, a preference hierarchy among the XAI methods, participants were asked the order of their most preferred XAI method to the least, with the first explanation (SHAP) method emerging as a clear favorite, the second method (Grad-CAM) falling significantly behind in preference, and the third method (LIME) landing somewhere in the middle.

Further difference in preference, some of which were previously touched upon, spanned a range of aspects but were not limited to: the desire for more in-depth explanations varied, with some participants seeking greater detail, while others preferred a more concise approach; and when it came to the format of explanations, opinions were split between those who favored visual representations over numerical data and those who valued a blend of both.

"I especially like explanation 1. It has the clearest ECG traces and there's also minimal marking, so it's not like you have a large blob in your face like the other two explanations. You just have the very localised area and the XAI telling you this is what I think this is going on. Everybody can get a marker and just make a big blur site roughly somewhere on the ECG. This (explanation 1) is very specific and that's what I like about it." - P2

"It took me some time to really understand the data for explanation 1 but for explanation 2, before I clicked open the the explanation, I can already see what's going on with this region. I like this the most because when I see it, I will directly go to that red mark position and check out what is wrong with that region." - P6

Showcasing the dynamic nature of individual preferences and the complexity of human cognition, one participant vividly illustrated how perceptions can shift even within a single interaction. Initially, this participant noted discrepancies in the ECG dataset and inconsistencies in the XAI's markings, which led to confusion. However, as the interview progressed, the same inconsistencies that initially puzzled him paradoxically increased his trust in the XAI system. They remarked:

"Like you can see that it's contradicting itself and I trusted it a little bit more for some reason, which is completely not scientific I know, but I liked the explanation more for it" - P3

These examples represent just a fraction of the diverse preferences expressed by participants, highlighting the complexity and individuality of user needs and expectations in XAI system design. This emergent principle emphasises the importance of tailoring explanations to meet the unique perspectives and preferences of each user.

5.5.9. XUI design

In addition to the previously discussed feedback, participants also offered specific suggestions regarding the design and layout of the XUI. They proposed modifications such as adjusting the placement of certain explanation methods to more preferred locations within the interface, however these were different per participant, and employing more contrasting colours for the XAI explanations against the background of the ECG traces, aiming for greater visual distinction and clarity.

"...the colour you have to watch out for, you could have a sharper contrast. Maybe try making this a darker or more intensive colour or something, make it stand out more. It feels like the red is on top of the image (ECG traces), while I would like to have the other way around the red behind the image." - P2

"...I think I would open this one (points to prediction score) by default with a big score to immediately see the certainty and maybe also this one (prominents leads) too, or in the middle as well with the red indication. But I would like to know what are the leads that were most predictive also in a different way, maybe in addition with the current explanation" - P4

Building on insights mentioned in earlier sections, there was a consensus among participants for more concise explanations that directly guide their focus to the relevant areas. The use of markings by the XAI methods, particularly when combined with explanations about prominent leads, was positively received. Such design elements that effectively direct users' attention to key points were highlighted as beneficial. Participants indicated that enhancements in this direction, facilitating a more focused guidance on where to look, would significantly improve their interaction with and understanding of the XUI.

"My eyes are drawn to both colours and I think the importance with MI is to detect immediately where which leads is indicating MI the most. So I think when you leave out the blue, your eyes only drawn to the red areas. It would be more straight to the point and that would be more helpful to me, I would have a better picture of what it was trying to explain to me. Because right now the negative blue marking is kind of confusing to me." - P7

6

Research Findings & Discussion

6.1. XUI design

Building on the insights gathered from thematic analysis, this section evaluates the selected design elements of the XUI.

6.1.1. Design principles

The application of design principles within the XUI was positively received by the participants, significantly enhancing the interpretability of the XAI methods employed. From the interview feedback, it was evident that the XUI played a crucial role in making the AI's decisions comprehensible. Participants noted that without the structured guidance of the XUI, they would have found it challenging to understand what the XAI output was indicating or even the rationale behind the AI's method, effectively demonstrating the importance of the user interface in aiding interpretation. Feedback for each applied design principle is presented in Table 6.1, re-iterated from section 4.3.2.

Table 6.1: Feedback on the implementation of the design principles

Complementary Naturalness	
Implementation	Impact
Textual explanations accompany visual representations, providing a verbal narrative that explains the underlying XAI methods in a way that complements the visual data.	Crucial—without it, participants would have been unable to understand what the XAI was depicting or the specifics of the XAI method.
A variety of textual descriptions are provided for different aspects of the XAI output, offering users multiple angles from which to interpret the output, including textual elaborations on the XAI method, the predictive score, and the prominent leads.	Very useful, not only for explaining the content of the explanation box itself—for instance, the prediction score alone would have been vague—but also for clarifying the output further.

Flexibility Through Multiple Ways To Explain	
Implementation	Impact
To offer a multifaceted perspective on diagnosis, the XUI implements three distinct XAI methods for analysing the same patient data, providing a flexible viewpoint.	A good balance in the number of options provided; more than three would have been overwhelming, and fewer could suffice but depends on the developmental phase of the XAI. Given that the use case is still exploratory, presenting multiple XAI outputs was appropriate.
The XUI combines visual and textual explanations, allowing users to see the XAI outputs and read about them in detail, aiding in the comprehension of the visual information.	As noted earlier, the combination of both elements was essential, complementing each other to provide a more comprehensive explanation.
The explanations include both numerical data and their textual interpretations. This is achieved by offering a textual explanation alongside a numeric score clarifying the significance of the score in plain language.	This element wasn't noted as particularly helpful, but also not deemed completely useless; its effectiveness in influencing how participants interpreted the numbers could not be definitively measured.
Responsiveness Through Progressive Disclosure	
Implementation	Impact
Elements within the XUI can be expanded or collapsed by the user, providing control over the flow and quantity of information consumed, which is particularly useful in managing cognitive load.	Useful; observations during the interviews revealed that each participant explored the XUI in their own unique way. Despite instructions and information provided beforehand to streamline the experience across different interviews, individual interactions varied significantly.
Some elements are set to a default closed position to streamline the user's focus and prevent information overload, while key information remains constantly visible to guide the user's attention to essential details.	Useful; despite the XAI being the largest visual element on the screen, the ability to close other boxes ensured that initial focus was on the XAI, reinforcing its central role. This design helped guide participants' attention back to the XAI as needed, especially after exploring other explanations, thereby solidifying its importance from the first interaction.

Determining the impact of individual design principles is complex due to their interconnected nature; it is difficult to isolate the effects of one without considering its interaction with others. For instance, the incorporation of textual explanations, an element of complementary naturalness, inherently involves offering multiple explanatory perspectives, which aligns with the principle of flexibility through multiple ways to explain. This integration naturally leads to considerations under the principle of responsiveness through progressive disclosure, where decisions about which explanations to disclose and in what order become essential. Although the design principle of sensitivity to context and mind was not explicitly implemented in the initial design, it emerged as a significant theme during the thematic analysis, highlighting its relevance and interconnection with other principles. The relationship among these principles and their collective impact on the XUI's effectiveness will be further discussed in section 6.2.

This section has addressed the influence of design principles on end-user interpretability; the subsequent section will examine how each individual explanation element within the XUI was received, discussing the feedback gathered to guide future enhancements and refinements of the XUI design.

6.1.2. Explanations elements

Patient information

Feedback on the patient information element was minimal, indicating a general consensus on its necessity and appropriateness. Participants universally appreciated the type of information displayed and the consistently disclosed nature of this element. This suggests that the current implementation effectively meets the users' expectations and needs, providing essential context without overwhelming the interface.

XAI method

This element provided a textual explanation for each of the XAI methods employed. The inclusion of textual explanations was identified as essential, particularly for users interacting with the system for the first time. The findings indicated that while this element might not be frequently used in regular interactions, it remains crucial for instances where users may require a refresher on the methods. Therefore, it is recommended that this element be retained in future designs, remaining accessible yet initially undisclosed. Furthermore, each explanation was crafted with varying levels of technical jargon to cater to different user proficiencies. Feedback on the preferred depth of these explanations was mixed, suggesting that no definitive conclusion could be drawn regarding the optimal level of detail. Thus, maintaining a balance in the complexity of information presented is advised in accommodating diverse user needs.

Prediction score

This element presented users with a numeric prediction score from the AI, accompanied by a textual explanation elucidating the significance of the score in plain language. This approach was informed by the literature, as noted in several studies (Josephson & Josephson, 1996; McClure, 2002; Miller, 2019), which suggest that solely relying on probabilities might not meet user satisfaction. Nevertheless, feedback from interviews indicated a preference for displaying the prediction score as a standalone element, without the accompanying textual explanation. While the text was considered useful for initial interactions, it was viewed as redundant for frequent users. The reliance on numerical probabilities and the necessity of textual explanations should be context-dependent, reflecting the specific needs and familiarity of the user with the system, thus determining the practical relevance and utility of this element in the XUI.

XAI output

This element provided users with textual explanations detailing the significance of the color markings and their impact on the XAI outputs. While necessary for initial interactions to help users understand the visual data, feedback indicated that this element was considered unnecessary for regular use, unlike other explanation elements. In future iterations of the XUI design, it is recommended to assess the specific application contexts of the XUI to determine whether such explanatory components should be retained or modified. This would ensure that the interface remains streamlined and user-centric, focusing on delivering relevant information as needed without overburdening experienced users.

Prominent leads

This element highlighted areas in the image identified by the XAI as most or least influential in the AI prediction, effectively guiding users' focus. Feedback indicated that while the markings for the most prominent leads were valuable for directing attention, the indicators for the least prominent leads were often disregarded as irrelevant. This mirrors responses to the SHAP method's negative markings, reinforcing the view that negative influences might not be crucial for MI diagnosis. Although seen as beneficial for initial interactions, for regular use, users preferred only the visual cues for prominent leads, suggesting that the accompanying descriptive text could be minimised or made collapsible to streamline the interface for experienced users.

Layout

The layout of the XUI could see improvements in context of frequent use. Synthesising the feedback on the explanation elements and XUI in general, the layout for frequent use should consider the following points :

- Maximising Visual Space for XAI Output: It's crucial to allocate the majority of the XUI's visual space to ECG traces/XAI outputs. Since these elements are central to the user's inspection and analysis, expanding their display area would facilitate a more detailed and thorough evaluation aiding in the interpretability of the output.
- Customisable Interface Options: User feedback varied significantly regarding the utility of different explanation elements. A customisable XUI would allow users to personalise the interface according to their specific needs and preferences, potentially increasing user satisfaction and interpretability. However, implementing such flexibility introduces practical challenges in design and user experience that need careful consideration.
- Optimising XAI Method Selection: While feedback on the number of XAI methods presented was mixed, there was a clear indication that reducing the options might simplify the user experience, whereas increasing the number would be too overwhelming. Allowing users to select their preferred XAI methods could offer a balanced solution, providing enough diversity in analysis without overwhelming the user.

The feedback on the XUI layout in context of first-time use varied widely, with no consensus on specific changes. Most participants found the current design and the amount of information provided satisfactory. While there was no strong preference for reducing information, some participants expressed a desire for more in-depth explanations and opportunities to explore the workings of the model further. However, such additions would depend on the XAI model used and might only require adjustments to accommodate more explanation boxes without shifting the focus away from the central XAI output, which should remain prominently displayed with explanatory elements surrounding it.

6.1.3. XAIs

Among the XAI methods evaluated, SHAP (XAI method 1) was the most preferred, followed by LIME (XAI method 3), and Grad-CAM (XAI method 2). Participants' preference for SHAP was primarily due to its clarity and precise guidance on critical points for analysis. Unlike Grad-CAM and LIME, which rely on image processing that can sometimes degrade image clarity, SHAP's approach maintains the quality of visual output, making it clearer for users to interpret. In terms of directing user attention, SHAP and LIME were preferred for their ability to guide users to specific areas or points, although opinions varied significantly among participants. Some found SHAP's detailed markings overwhelming or confusing due to the extensive use of blue to denote negative importance. Conversely, those who favored LIME appreciated its focused approach on specific regions over pinpointed entries, finding it less overwhelming.

The evaluation suggests several areas for improvement and consideration in future XAI development for MI diagnosis:

- Clarity and Direction: The primary feedback points to the need for clearer explanations and more precise guidance on where to look, which are crucial for effective user interaction with XAI outputs.
- Suitability of Methods: The appropriateness of using Grad-CAM and LIME for MI detection is questioned based on current feedback, suggesting a potential reevaluation of the methods or the way visual outputs are generated. For instance, superimposing LIME's heatmap over a clearer image might enhance its utility in medical diagnosis.
- Relevance of Negative Markings: The use of negative markings in SHAP, while informative, was sometimes found to be distracting or irrelevant for diagnosing MI, suggesting a need to tailor visual cues more closely to the specific diagnostic requirements of medical use cases.
- Exploration of colour usage in the XAI output could be further refined. Participant feedback highlighted the need for contrasting colours to distinguish between the XAI output and the underlying ECG traces more clearly. This suggests that enhancing visual differentiation could improve user comprehension and interaction with the system, making it easier to differentiate between AI-generated insights and actual ECG data.
- Consideration should be given to arranging ECG traces in a conventional layout to enhance readability. Additionally, the scalability of ECG traces should be considered when generating the XAI output, as participants expressed a desire for the ability to zoom in for closer inspection of specific details within the image.

The improvement points above can be extended to the discussion as highlighted in section 6.4, which revolves around aligning XAI outputs with clinical expectations versus explaining them from an AI model's perspective. This discussion is pivotal in determining the direction of future XAI designs, whether they should conform to standard clinical diagnostic practices or provide insights based on the underlying AI models' operations.

6.1.4. Deployment of XUI

Understanding the deployment context of an XAI is crucial in determining its design and interaction dynamics. The distinction between the intended use and practical application of the XAI highlights how user interaction can vary significantly based on the context—whether for educational purposes, initial familiarisation, or clinical decision-making.

Results from user feedback reveal distinct preferences and requirements depending on the imagined use-case scenario. For educational or introductory contexts, users may prefer more detailed explanations and interactive features to aid in understanding the model's workings. Conversely, in a clinical setting where decision-making is paramount, the emphasis shifts towards efficiency, clarity, and quick access to relevant information.

Recognising these differing needs underscores the importance of adaptable and context-sensitive design approaches in both XAI and XUI development. This flexibility not only caters to the varied educational backgrounds of the users but also accommodates the urgency and accuracy required in clinical environments, thereby enhancing the overall interpretability and effectiveness of the XAI system in real-world applications.

6.2. Proposition to Design Principles

The three interactive design principles—complementary naturalness, responsiveness through progressive disclosure, and flexibility through multiple ways to explain—were incorporated into the XUI and examined through interviews. During these interviews and subsequent code analysis, a consistent theme emerged across these principles: the pivotal role of context. In the literature review, key design principles were identified for the design of the XUI based on the paper by (Chromik & Butz, 2021). In the paper, the importance of context and individual preferences within each principle are recognised, highlighting how these elements are integral to the design principles :

- **Complementary naturalness** through its variety of visual and textual explanations, directly caters to the diverse contexts and cognitive preferences of users. By incorporating contextual cues and acknowledging individual user preferences, this principle accommodates explanations that are tailored to fit the unique mental models and learning styles of each user.
- **Responsiveness through progressive disclosure** adapts the depth of explanation to match users' specific contextual needs and personal preferences. It gives users the flexibility to explore information at a pace that suits them, catering to different levels of prior knowledge and individual cognitive styles for grasping complex concepts.
- **Flexibility through multiple ways to explain** recognises the diversity of users' mental models and situational contexts. This principle enables users to choose explanation methods that best fit their cognitive frameworks and meet their specific needs, fostering a personalised understanding that is finely tuned to each individual's preferences.

Despite these recognitions, the original approach treats all four interactive design principles, including "Sensitivity to the mind and context," with equal emphasis. This approach does not elevate any specific principle as foundational over the others, despite the clear influence of context and personal preferences noted in their discussions. This conventional approach overlooks the distinct impact that an in-depth understanding of context and individual cognitive processes exerts on the interpretability of XAIs.

The concept of structuring these interactive design principles into a hierarchical framework was derived from insights gathered during the interviews. Participants displayed a wide array of preferences, illuminating the diversity in how users engage with and interpret the XUI. Notably, despite the fourth principle of sensitivity to mind and context not being explicitly integrated into the XUI, its importance naturally

surfaced through participants' feedback, accentuating its fundamental role above the other principles in the context of XUI design.

By suggesting a pyramid ranking with "Sensitivity to the mind and context" at the pinnacle, this framework acknowledges that understanding and adapting to the individual user's context and preferences is foundational. Positioned beneath this top principle, complementary naturalness, responsiveness through progressive disclosure, and flexibility through multiple ways to explain act as critical supports. These principles are vital for crafting an interpretable XUI, but their effectiveness is substantially amplified when they are informed by a comprehensive understanding of the user's specific needs and cognitive styles.

As discussed in the previous section on the effectiveness of the applied design principles within the XUI, these design principles are intrinsically interconnected, including "sensitivity to mind and context." This interconnectedness is visually represented in Figure 6.1, where all four principles are depicted as part of a pyramid structure. The three principles at the base of the pyramid—complementary naturalness, responsiveness through progressive disclosure, and flexibility through multiple ways to explain—are intertwined in a horizontal manner, meaning their implementation indirectly influences one another. In contrast, the principle of "sensitivity to mind and context" at the apex of the pyramid has a more vertical and direct relationship with the other three. This vertical connection means that any application of the top principle directly affects the implementation of the principles at the base of the pyramid. However, due to the horizontal linkage among the principles at the bottom, the influence exerted by the top principle is uniformly distributed across them. Therefore, while "sensitivity to mind and context" directly enhances how each foundational principle is applied, it does not disproportionately affect any single principle due to the balanced, horizontal interconnections among them.

This nuanced approach to organising interactive design principles for XUIs, inspired by the compilation of interview feedback, suggests a more targeted methodology for XUI design. By prioritising the adaptation to individual differences and contexts, designers can create XUIs that are not only more intuitive and centered around the user but also highly responsive to the varied interpretive needs of the audience. This notion that the explanations should be tailored to the mental model of the user was mentioned as one of the factors of interpretability under section 3.1.8 "Explanation Evaluation" (Read & Marcus-Newhall, 1993; Thagard, 1978). This notion is also supported by other work in XAI mentioned in the literature review (D. Wang et al., 2019). Adopting this strategy has the potential to markedly improve the interpretability of the XAI, ensuring that interactive design principles play a pivotal role in developing XUIs that are both accessible and user-friendly.

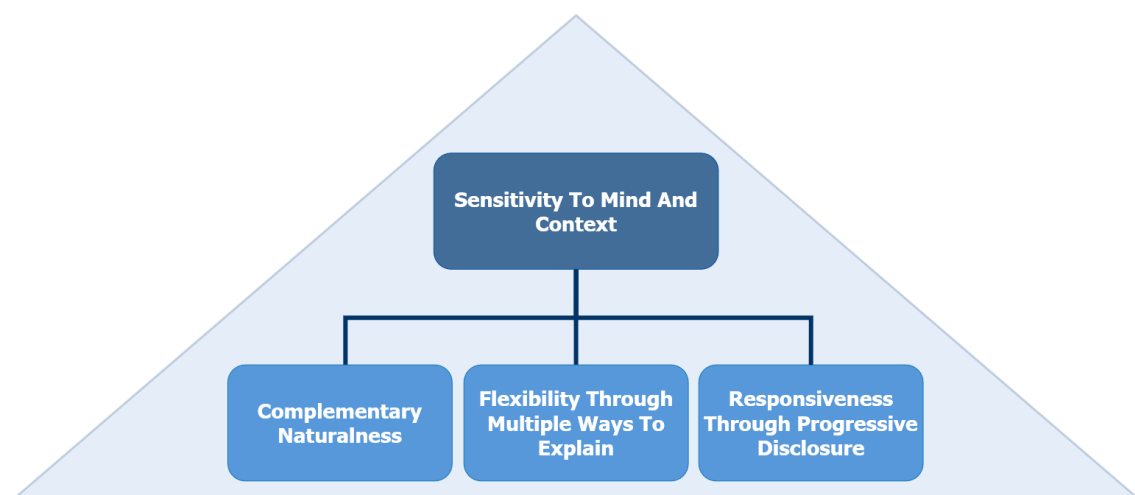


Figure 6.1: Proposed restructuring of the Design Principles by Chromik and Butz (2021)

6.3. Conceptual framework reflection

The interview results reveal that the initial communication gap between the FOKUS XAI designers and their target audience triggered a domino effect on subsequent parts of the conceptual framework. Research indicated a disconnect between the target group, professionals skilled in reading ECGs for acute MI, the domain context of the ECG data used, and the XAI designers. This misalignment adversely affected the XAI outputs, which did not meet the expectations of the target group. Participants noted problems stemming from the use of an incorrect ECG dataset and suboptimal visualisation of ECG layouts and traces. Although some participants were able to partially overlook these issues, the overall interpretability of the XUI was compromised, as evidenced in the findings. This domino effect is illustrated in Figure 6.2 with red-coloured connections indicating the problematic areas. The dashed arrow represents the disconnection between the target group/domain context and the XAI outputs, leading directly to challenges with the types of interactive elements available, as depicted by a solid red arrow pointing towards this segment. This sequence ultimately impacted the interpretability, as shown in the subsequent block of the diagram.

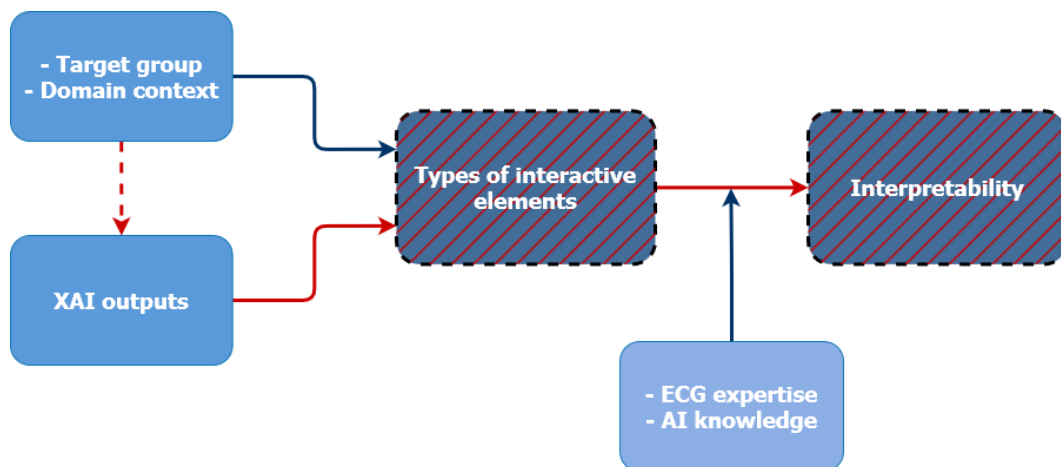


Figure 6.2: Issues identified in the conceptual framework

Regarding the impact of ECG expertise and AI knowledge on interpretability, the findings from the interviews were ambiguous. Although it was anticipated that greater expertise in these domains might enhance interpretability, no significant differences were observed between participants with varying levels of expertise. This suggests that individual preferences and mental models significantly influence interpretability, which was not conclusively impacted by the participants' depth of knowledge in ECG or AI. This inconclusive outcome, coupled with the prototype's limitations due to scope and time constraints, indicates that further research is necessary to fully explore how expertise in ECG and AI could influence interpretability in XAI systems.

Reflecting on FOKUS' emphasis on exploring XAI design methodologies rather than perfecting explanations underscores a critical oversight in their approach: the delayed involvement of the target group. This decision adversely affected their XAI models, demonstrating that the initial design strategy was inefficient. The lack of alignment between the XAI's development and the specific needs and context of the target audience compromised the effectiveness of the entire project. This situation highlights the crucial need for clear, effective communication from the outset, ensuring a deep understanding of the target users' requirements to forge a more interpretable and effective XAI system. Proactive engagement with the target audience is essential to tailor development to their specific needs, thereby enhancing the XAI system's utility and interpretability. A solid foundation in XAI development is fundamental in crafting an effective XUI, showing that foundational aspects cannot be ignored. Thus, involving the target group is vital not only for XUI development but also for advancing XAI methodologies, illustrating that exploring design methodologies cannot be isolated from refining explanations.

Conceptualising the findings from the research and reflecting on the conceptual framework led to the development of a new design approach, depicted in Figure 6.3. Although the scope of this thesis did not include the direct development of XAI, the research underscored that XAI development is a critical

phase preceding XUI development. This phase is vital for ensuring that the XUI effectively communicates and aligns with user needs and expectations. Consequently, the proposed design approach incorporates suggested steps for XAI development to streamline the subsequent XUI development process.

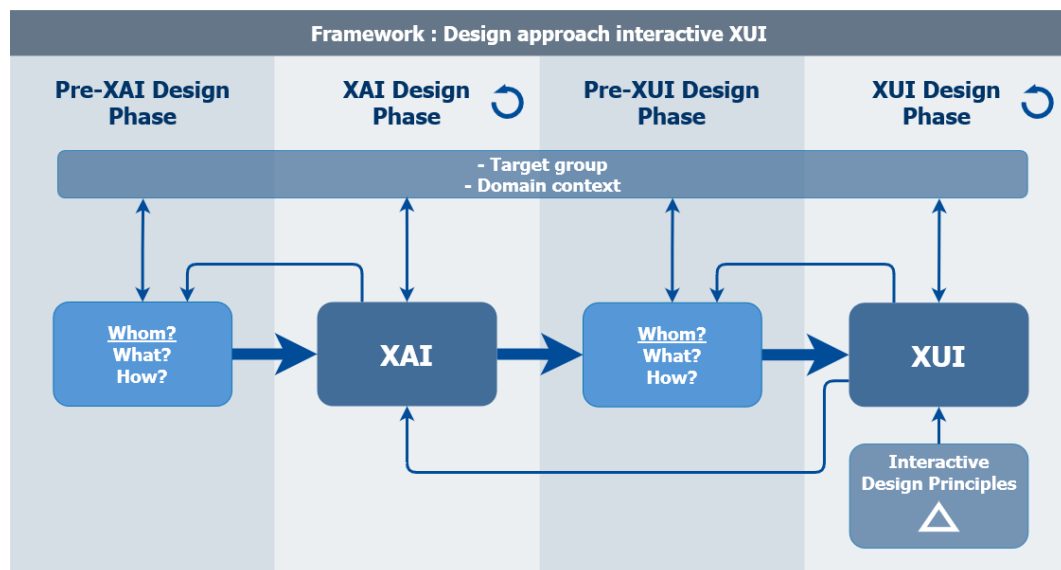


Figure 6.3: Framework: Design approach for interactive XUI

The design approach to interactive XUIs is structured into four sequential phases: pre-XAI design, XAI design, pre-XUI design, and XUI design. The initial phase, pre-XAI design, sets the foundation by addressing three critical questions identified by Miller et al. (2017): whom to explain to, what to explain, and how to explain. This sequence effectively establishes the scope of the XAI by prioritising the identification of the target group and domain context, which then informs the approach to the subsequent questions, such as choosing the appropriate explanation model suited for the use case. This approach is mirrored in the pre-XUI design phase, underscoring its effectiveness in delineating scope.

In the XAI design phase, the specifics of XAI development are not fully defined within this framework, but it involves continuous engagement with the target group and domain context, iterating as necessary to refine or expand the scope defined in the previous phase. The iterative nature of this phase is symbolised by a circular icon alongside the phase title in the framework.

Once the XAI has matured to a stage suitable for integration with an XUI, the pre-XUI design phase revisits the initial questions, tailoring them to the developed XAI methods and how these can be explained with the same target group and domain context in mind. This phase ensures the scope for the XUI is clearly defined based on the outputs from the XAI phase.

In the fourth and final phase of this design approach, the focus shifts to the actual development of the XUI. This phase benefits significantly from the clearly defined scope established in the preceding phase, providing a solid foundation for applying design principles effectively. Communication remains crucial in this phase, particularly with the target group and domain context, to ensure that the XUI aligns closely with the users' mental models and contextual needs. This alignment is vital for effectively applying the principle of sensitivity to context and mind, which cannot be optimised without ongoing feedback and collaboration.

The practical steps in this phase start by reviewing the possible XAI implementations outlined in the scope, giving designers a comprehensive overview of the explanatory options available. A critical activity in this phase is the selection of specific explanations to implement, a decision that may involve collaboration with the target group to ensure relevance and appropriateness. Following this selection, the implementation of these explanations is guided by the core design principles, with a continuous process of cross-referencing to ensure that each implementation maximises effectiveness.

After the initial implementation, it is essential to collect feedback from the target group. This feedback is instrumental in enhancing the customisability of the XUI, focusing particularly on optimising the top design principle—sensitivity to context and mind—to ensure the XUI offers maximum interpretability and usability. The degree of customisation will depend on project constraints and practical feasibility within the given timeline.

Like the XAI design phase, the XUI design phase is inherently iterative. Depending on the outcomes and feedback, there may be a need to revisit earlier phases to refine the scope or even to step back to the XAI design phase if significant adjustments are required. This iterative process ensures that the final product not only adheres to technical specifications but also resonates effectively with the end-users, thereby enhancing the overall effectiveness of the XUI.

Throughout these phases, the framework emphasises the necessity of iterative design and constant communication with the target group. This approach not only aligns the XAI and XUI with the users' needs but also ensures that the interactive design principles are effectively applied to enhance the overall interpretability and functionality of the system. The goal is to create XAI and XUI that are not only technically sound but also intuitively aligned with the end-users' expectations and requirements.

6.4. XAI design

This research primarily examined the influence of XUI design principles on the interpretability of XAI. However, it also unearthed several insightful aspects concerning XAI and AI design. Apart from the use of an incorrect dataset, an intriguing issue was the marking of ECG segments by XAI that diverged from conventional medical training. Initially perceived as a problem, this discrepancy prompts a reconsideration of whether such a divergence is indeed a flaw or a fundamental feature of XAI. The core aim of XAI is to demystify the decision-making processes of AI, especially in complex deep learning models that are often opaque "black boxes."

This raises a critical question: Is it problematic for XAI explanations to diverge from established medical diagnoses? XAI's primary function is to elucidate the AI model's decision-making process rather than to mimic standard ECG interpretations. However, in scenarios where XAI supports decision-making, it is crucial to assess whether explanations that deviate from a user's expectations or understanding could compromise the tool's effectiveness. If such explanations lead to confusion, the utility of XAI as a decision-support tool may be undermined.

Moreover, this situation highlights a broader issue in the application of XAI: the balance between full transparency and practical usability. While full transparency is invaluable for diagnostic or auditing purposes, in decision-support contexts, fidelity to an AI's internal reasoning might be less critical than providing explanations that align with user mental models. Thus, for decision-support applications, focusing on reinterpreted or intermediate explanations that resonate more closely with how users think and operate might enhance the practical effectiveness of XAI.

Considering the original goals of the use case XAI, employing it in a clinical setting, as in the FOKUS use case, might not be entirely suitable due to the potential for confusion. Modifying the XAI to align more closely with expert perspectives might, in fact, stray from the fundamental objective of XAI, which is to illuminate the AI's rationale rather than conform to user expectations based on their familiarity with the subject matter.

This leads to a pivotal question for the future direction of XAI research: Should XAI focus on offering explanations that accurately reflect the AI's reasoning process, or should it aim to provide explanations that align with user expectations, particularly in areas where they possess expertise? This discussion underscores the importance of a balanced approach in the development and application of XAI. It suggests that XAI research should prioritise both objectives, as the ultimate goal of XAI may vary depending on the deployment context. Therefore, the design patterns adopted should also differ, tailoring to the specific needs and expectations based on the deployment environment. This approach would respect the technology's intent while remaining adaptable to user needs and expectations, ensuring that XAI remains a valuable tool across various applications. This concept aligns with the pyramid model for design principles, emphasising the fundamental role of user customisation, and the proposed XUI design framework that advocates for iterative design and ongoing dialogue as crucial, ensuring that XUIs are both interpretable and effective.

7

Conclusion

This research focused on the design and implementation of interactive design principles of an XUI to enhance the interpretability of XAI for end-users. This chapter answers the main research question related to the identified knowledge gap, summarising the study's findings. It concludes by discussing the research's limitations and outlining potential directions for future work in this area.

7.1. Layered Knowledge Gap

Two primary knowledge gaps were identified in XAI interpretability. The first concerns the fact that while current explanation methods employ various strategies to enhance interpretability, they primarily cater to users with a technical background, often neglecting end-users who lack such expertise. This oversight emphasises the need for explanation strategies that are consciously designed with the end-user in mind, ensuring that interpretations are accessible and comprehensible across a broad spectrum of understanding. The second gap involves the practical application and development of user-centric design principles for XUIs. Despite theoretical acknowledgment of these principles' importance, their real-world application, especially in environments where actual users interact with AI, remains under-explored. This knowledge gap is addressed by answering the main research question in the following section.

7.2. Main research question

How can interactive design principles be strategically applied to Explainable User Interfaces (XUI) to enhance the interpretability of Explainable Artificial Intelligence (XAI) for end-users?

This research employed the Design Science Research Method to navigate the intricate challenge posed by the main research question. An in-depth literature review was initially conducted, examining critical areas such as human interpretability factors, human-computer interaction dynamics, and design principles for XUIs proposed by Chromik and Butz (2021).

The case study of the XAI project at FOKUS was selected for its real-world application, highlighting not just the current challenges and dynamics within the XAI field but also the taken approach of the FOKUS team in a domain unfamiliar to them. FOKUS's project, dedicated to real-time prediction of myocardial infarctions using ECG data, aims to use their XAI to explain the decision-making processes of the underlying models through visual representations.

Building on the literature review and the FOKUS use case, a conceptual and theoretical framework was devised. The theoretical framework incorporated insights into interpretability and human-computer

interaction, aligning them with three out of four key design principles by Chromik and Butz (2021) : complementary naturalness, responsiveness through progressive disclosure, and the flexibility of multiple ways to explain. These principles served as the foundation for the XUI design, guiding its development to enhance end-user interpretability.

The research proceeded to validate the conceptual framework through the practical application of the XUI developed for FOKUS's XAI, supplemented by semi-structured interviews with (medical) experts versed in cardiology. Thematic analysis of the interview data revealed challenges within the XAI that affected its interpretability, notably due to dataset misuse and visualisation errors, pointing to an oversight in stakeholder engagement by the FOKUS XAI team. Despite these hurdles, the core design principles demonstrated their effectiveness in enhancing XAI interpretability, particularly the 'Sensitivity to the mind and context' principle. Though not fully realised in the initial XUI iteration, this principle was deemed crucial for its acknowledgment of the diverse interpretive lenses and preferences among users.

The research proposes organising these design principles in a pyramid model, with "Sensitivity to context and mind" at the top. This model suggests that while all principles are valuable, they should be applied with a primary focus on context and user needs. Ideally, an interface would be adaptable, capable of being customised to fit anyone's preferences, reinforcing the notion that there's no one-size-fits-all solution. Still, it is recognised that this would be hard to achieve in practice. Alongside this restructuring, a framework for a design approach to XUI was proposed, which conceptualises the findings of the research and the evaluation of the initial conceptual framework. The integration of these two propositions aims not only to align the XAI and XUI with users' needs but also to ensure that the interactive design principles are effectively applied, enhancing the overall interpretability and functionality of the system.

The biggest takeaway from the research is the crucial role of context in creating effective and interpretable XUIs. Engaging with stakeholders might appear to be a straightforward strategy, but it requires a nuanced approach to XUI design, taking into account the specific context and demands of the project. The findings of the research highlight the necessity for more in-depth research into creating interpretable XUIs. The exploration of Chromik and Butz (2021) design principles within an operational XUI marks a first in the field, suggesting potential enhancements to improve XAI interpretability further. This blend of insights from the research offers a new perspective on adapting and applying design principles to meet the evolving needs of interpretability in XAI.

7.3. Limitations and future research

Interview Design and Execution

The researchers limited experience with conducting in-depth interviews had some effect on the design and execution of the interview process. Some questions may have inadvertently influenced participants' responses, while others did not capture the full scope of the research interests. This occasionally necessitated follow-up questions during the interviews to clarify initial responses, indicating room for improvement in both the design and facilitation of the interviews. Future research could benefit from more structured training in interview techniques and the development of a more robust interview protocol. These steps would help ensure that the questions are well-crafted and effectively capture the full scope of the research interests, leading to more reliable data collection.

Interplay Between Model Understanding and Decision-Making

One of the central challenges was attempting to simultaneously enhance users' understanding of the AI model and facilitate expedited decision-making. The project endeavored to address both these aspects; however, due to inherent limitations, neither objective was fully achieved. This dichotomy resulted in a scenario where the project could not optimally leverage the potential benefits of deep model understanding to significantly improve decision-making efficiency. Future projects could benefit from developing targeted strategies that separately address model understanding and decision-making speed, possibly through adaptive interfaces that adjust based on the user's familiarity and interaction context.

Exploration of (X)AI in the Medical Domain

The goal of exploring the application of (X)AI within the nuanced and complex medical field was constrained by several factors, including the project's tight timelines and the intricate nature of medical (X)AI itself. This led to a surface-level examination that could not delve into the nuanced implications of current medical (X)AI advancements on the research and the design of the XUI. A deeper investigation into these aspects could potentially provide valuable insights into how (X)AI could be more effectively tailored for medical applications. Engaging in deeper collaborations with medical experts could enrich the research, providing nuanced insights into integrating (X)AI effectively within medical practices.

Impact of XAI Development Stage on Research

The limitations inherent in the developmental stage of the XAI system had a profound effect on the entire research process, from the formulation of interview questions to the analysis of participant feedback. These shortcomings meant that feedback often pertained to broad issues rather than specific actionable design improvements. A more mature XAI system might have facilitated a more focused exploration of design elements that could be optimised. Collaborating with domain experts from the outset could ensure that the XAI development is aligned with end-user needs and the practical requirements of specific domains. This collaboration should carefully consider whether to prioritise explanations that reflect the AI's reasoning process or those that align with user expectations, based on insights from both the XAI field and domain-specific knowledge. Future work could utilise the proposed design approach framework as a guideline to effectively engage stakeholders during the crucial development stages of both XAI and XUI design.

Absence of Iterative XUI Design Revisions

Despite acknowledging the importance of the feedback received in the initial round of interviews, the project did not undertake revisions to the XUI based on this feedback. This decision was influenced by both the recognition of the challenges associated with integrating the suggested design principles and the practical limitations of the project's scope and timeline. The inability to test the implementation of these principles in practice represented a missed opportunity to iteratively refine the XUI based on user insights. Implementing a more agile and iterative design process in future projects could allow for the continuous integration of user feedback into the XUI design, as proposed by the design approach framework. The initial design of the XUI could be considered as the foundational step in an iterative process. The feedback gathered during the initial evaluations can serve as a valuable guide for subsequent iterations. Future research should aim to utilise the insights from this initial feedback to inform the development of the next iteration of the XUI. This approach would enable ongoing refinement and adaptation of the XUI to better meet user needs and expectations, ultimately enhancing the utility and relevance of the XUI in varied decision-support contexts.

7.4. Reflections

Scientific contributions

This research substantially enriches the academic landscape by practically applying and refining the design principles outlined by Chromik and Butz (2021), demonstrating their implementation in real-world scenarios for non-technical end-users. It introduced a novel restructuring of these principles into a pyramid model, enhancing the effectiveness of XUI implementations to improve system interpretability. Additionally, the study developed a comprehensive design approach framework for XUI that incorporates XAI design as an integral phase, emphasising the importance of aligning XAI development with user needs to ensure effective XUI outcomes. This framework not only demonstrates a practical application of theoretical design principles for non-technical end-users but also provides XUI designers with actionable guidelines and a clear methodological approach to optimise user interpretability in XAI systems. Moreover, the research extends beyond the confines of XUI development to impact the broader field of XAI by illustrating the consequential effects of XAI design decisions on user interpretability and system functionality. This dual contribution addresses existing gaps in both XUI and XAI research, advancing the practical deployment and understanding of AI technologies across various user groups, thereby fostering broader acceptance and ethical integration of these systems into everyday applications.

Societal contributions

This research has made propositions to enhance the interpretability of AI systems. By making complex AI processes more comprehensible and accessible, it fosters trust and engagement among the general public, which is crucial as AI technologies increasingly permeate various societal domains (Doshi-Velez & Kim, 2017; Jin et al., 2021). In addition, fostering this trust and engagement supports the ethical integration of AI by addressing key concerns such as bias elimination and the promotion of fairness and inclusivity (Gunning & Aha, 2019).

Moreover, this research underpins a more informed public discourse on AI, as it was targeting end-users, facilitating discussions on its ethical and practical implications. By improving how AI systems are interpreted, the study aids in aligning AI advancements with public interests and regulatory frameworks, ensuring that technological progress does not outpace ethical considerations. The implementation of such research fosters a safer, more just integration of AI into society, promoting the well-being and welfare of all stakeholders involved (Doshi-Velez & Kim, 2017; Rajpurkar et al., 2022)

Management of Technology (MOT) Programme

This research began as an initiative to slightly diverge from the conventional curriculum of the Management of Technology (MOT) program and was driven by a personal desire to undertake more tangible research as opposed to the predominantly non-tangible research topics typical of the Master's program. Despite this shift in approach, the thesis topic and its findings are still intricately linked with the core subjects of the program. At first glance, the main focus on enhancing interpretability might not seem directly related, but the overarching theme aligns closely with the fundamental principles of the MOT program.

As highlighted in both the scientific and societal relevance sections, the adoption of AI technologies is gaining traction but faces challenges due to regulatory constraints and the current state of research, which still lacks effective methods for implementing such technologies. The proposed design approach framework and the pyramid model of design principles are efforts to bridge this gap and facilitate the adoption process of AI across various sectors. This directly addresses one of the crucial elements of the MOT program: how firms can leverage technology to increase productivity, profitability, and competitiveness.

Extending this further, the thesis directly engages with several key questions that the MOT curriculum aims to address, such as determining the necessary technologies for firms and the best strategies for acquiring these technologies—whether through internal development, collaboration, or external acquisition. A significant theme covered in every MOT course is the appropriate engagement of stakeholders. The findings of this thesis underscore the importance of this aspect and demonstrate how timely and active engagement can significantly propel technological development. The research sheds light on how these technologies can be effectively integrated into company operations and strategies, a point especially relevant given the program's focus on preparing students to manage and implement technology in alignment with organizational goals and market pressures.

Lastly, the courses from the master's program equipped the researcher with the skills to analyse qualitative data, conduct interviews, and think creatively to develop new recommendations and ideas. This background was crucial in enabling a thorough and innovative approach to the research, further bridging the gap between academic theory and practical application in the field of technology management.

Personal

Overall, I am happy with the thesis I was able to complete and deliver. Pursuing a topic that not only deviated slightly from the standard program but also delved into the intricate and nuanced interplay of various fields was definitely a bit risky. However, this challenge made the process personally enjoyable, albeit a bit frustrating at times. It provided me with invaluable insights into research methodologies as well as topics in (X)AI, social sciences, and HCI.

The journey of my thesis was definitely filled with a lot of bumps. Recruiting interview participants proved to be very difficult, as I was trying to engage medical professionals who were understandably preoccupied with treating patients and saving lives. Nevertheless, I managed to recruit eight participants who brought a diverse range of backgrounds to the study. Everyone was enthusiastic and provided helpful feedback, which was incredibly supportive and made the interview process smoother. This being my

first experience conducting interviews, I noticed an improvement in my ability to direct the interviews and interject questions as needed to delve deeper. This was a shift from my initial attempts, where I tried to stick to the protocol more closely and potentially overlooked valuable areas of discussion.

I am also happy to have collaborated with the FOKUS team. Although there was some initial confusion, once we established a clear scope and direction, our discussion sessions turned out to be both enjoyable and fruitful, resulting in interesting design ideas for the XUI. Had there been more time for this thesis, I would have implemented a revised iteration of the XUI, also incorporating changes from FOKUS' side to their XAI, which would have made this research even more substantial.

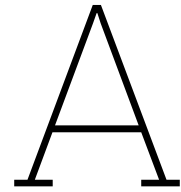
Looking back, I would have liked to engage in more research activities within the thesis, as the topic involved so many interesting insights across different fields. I have only just scratched the surface—really, it's more like dipping my toes into the ocean—of human understanding and interface design. I was quite overwhelmed with the volume of information I managed to gather and the challenge of applying it. Fully exploring all the nuances of the insights gathered would have been too much for the scope of this master's thesis and would likely warrant a full PhD research project. Having that said, I am satisfied with the knowledge I have acquired through this thesis, and I hope it will inspire future work in the direction of XUI and XAI design.

References

- A., S., & R., S. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7, 100230. <https://doi.org/https://doi.org/10.1016/j.dajour.2023.100230>
- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–18. <https://doi.org/10.1145/3173574.3174156>
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE access*, 6, 52138–52160.
- Ahmad, J., Farman, H., & Jan, Z. (2019). Deep learning methods and applications. In *Deep learning: Convergence to big data analytics* (pp. 31–42). Springer Singapore. https://doi.org/10.1007/978-981-13-3459-7_3
- Alfwzan, W. F., Alballa, T., Al-Dayel, I. A., & Selim, M. M. (2024). Statistical similarity matching and filtering for clinical image retrieval by machine learning approach. *Physica Scripta*, 99(1), 015020. <https://doi.org/10.1088/1402-4896/ad1668>
- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99, 101805. <https://doi.org/10.1016/j.inffus.2023.101805>
- Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. *Appl. Sci.*, 11(11), 5088. <https://doi.org/10.3390/app11115088>
- Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Ronny Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, Karthikeyan, Singh, M., Varshney, K. R., ... Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. <https://arxiv.org/abs/1909.03012>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 82–115. <https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012>
- Bethel, C. (2009). Robots without faces: Non-verbal social human-robot interaction.
- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2019). Explainable machine learning in deployment. *CoRR*, <abs/1909.06342>. <http://arxiv.org/abs/1909.06342>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., & Parikh, D. (2017). It takes two to tango: Towards theory of ai's mind.
- Chromik, M., & Butz, A. (2021). Human-xai interaction: A review and design principles for explanation user interfaces. *IFIP Conference on Human-Computer Interaction*, 619–640.
- Comission, E. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence.
- De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). *2017 AAAI Fall Symposium Series*.
- Dijk, O. (n.d.). Explainerdashboard. <http://titanicexplainer.herokuapp.com/>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning [cite arxiv:1702.08608]. <http://arxiv.org/abs/1702.08608>
- FOKUS, F. (20223). The fraunhofer institute for open communication systems. <https://www.fokus.fraunhofer.de/en>

- Gunning, D., & Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. *AI magazine*, 40(2), 44–58.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). Xai - explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>
- Hampton, J., & Hampton, J. (2019). *The ecg made easy e-book: The ecg made easy e-book*. Elsevier Health Sciences.
- Harman, G. H. (1965). The inference to the best explanation. *The philosophical review*, 74(1), 88–95.
- Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review*, 93(1), 75.
- Hilton, D., McClure, J., & Ben Slugoski, R. (2005). The course of events: Counterfactuals, causal sequences and explanation. *The Psychology of Counterfactual Thinking*.
- Holzinger, A., & Muller, H. (2021). Toward human–ai interfaces to support explainability and causability in medical ai. *Computer*, 54(10), 78–86.
- Hornbæk, K., & Oulasvirta, A. What is interaction? In: *Chi '17: Proceedings of the 2017 chi conference on human factors in computing systems*. New York, NY: Association for Computing Machinery, 2017, May, 5040–5052.
- Hulsen, T. (2023). Explainable artificial intelligence (xai): Concepts and challenges in healthcare. *AI*. <https://api.semanticscholar.org/CorpusID:260859950>
- Jin, W., Carpendale, S., Hamarneh, G., & Gromala, D. (2019). Bridging ai developers and end users: An end-user-centred explainable ai taxonomy and visual vocabularies. *Proceedings of the IEEE Visualization, Vancouver, BC, Canada*, 20–25.
- Jin, W., Fan, J., Gromala, D., Pasquier, P., & Hamarneh, G. (2021). Euca: A practical prototyping framework towards end-user-centered explainable artificial intelligence. *arXiv preprint arXiv:2102.02437*.
- Josephson, J. R., & Josephson, S. G. (1996). *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.
- Keil, F. C. (2006). Explanation and understanding. *Annu. Rev. Psychol.*, 57, 227–254.
- Leddo, J., Abelson, R. P., & Gross, P. H. (1984). Conjunctive explanations: When two reasons are better than one. *Journal of Personality and Social Psychology*, 47. <https://doi.org/10.1037/0022-3514.47.5.933>
- Lewis, D. (2013). *Counterfactuals*. John Wiley & Sons.
- Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27, 247–266.
- Loh, H., Ooi, C., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 226. <https://doi.org/10.1016/j.cmpb.2022.107161>
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3), 232–257.
- Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology*, 61(4), 303–332.
- Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions.
- Madumal, P., Singh, R. R., Newn, J., & Vetere, F. (2018). Interaction design for explainable AI: workshop proceedings. *CoRR*, abs/1812.08597. <http://arxiv.org/abs/1812.08597>
- McClure, J. (2002). Goal-based explanations of actions and outcomes. 12(1), 201–235.
- McClure, J., & Hilton, D. (1997). For you can't always get what you want: When preconditions are better explanations than goals. *British Journal of Social Psychology*, 36. <https://doi.org/10.1111/j.2044-8309.1997.tb01129.x>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1–38.
- Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences.
- Müller, J., Cypko, M., Oeser, A., Stoehr, M., Zebralla, V., Schreiber, S., Wiegand, S., Dietz, A., & Oeltze-Jafra, S. (2021). Visual Assistance in Clinical Decision Support. In S. Oeltze-Jafra & R. G. Raidou (Eds.), *Eurovis 2021 - dirk bartz prize*. The Eurographics Association. <https://doi.org/10.2312/evm.20211075>
- O'Sullivan, S., Janssen, M., Holzinger, A., Nevejans, N., Eminaga, O., Meyer, C., & Miernik, A. (2022). Explainable artificial intelligence (xai): Closing the gap between image analysis and navigation

- in complex invasive diagnostic procedures. *World Journal of Urology*, 40, 1–10. <https://doi.org/10.1007/s00345-022-03930-7>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45–77.
- Prasada, S. (2017). The scope of formal explanation. *Psychonomic bulletin & review*, 24, 1478–1487.
- Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, 99(1), 73–112.
- Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022, January). Ai in health and medicine. <https://www.nature.com/articles/s41591-021-01614-0#citeas>
- Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3), 429.
- 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>
- Robinson, O. (2014). Sampling in interview-based qualitative research: A theoretical and practical guide. *Qualitative Research in Psychology*, 11. <https://doi.org/10.1080/14780887.2013.801543>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2), 336–359. <https://doi.org/10.1007/s11263-019-01228-7>
- Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems*, 10(26), 1–31. <https://doi.org/10.1145/3419764>
- Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *The journal of philosophy*, 75(2), 76–92.
- Tintarev, N. Explanations of recommendations. In: *In Recsys '07: Proceedings of the 2007 acm conference on recommender systems*. New York, NY: Association for Computing Machinery, 2007, October, 203–206.
- Tsai, C.-H., & Brusilovsky, P. Evaluating visual explanations for similarity-based recommendations: User perception and performance. In: *In Umap '19: Proceedings of the 27th acm conference on user modeling, adaptation and personalization*. New York, NY: Association for Computing Machinery, 2019, June, 22–30.
- van der Velden, B. H., Kuijff, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79, 102470. <https://doi.org/10.1016/j.media.2022.102470>
- Wagner, P., Strodthoff, N., Boussejot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W., & Schaeffter, T. (2020). Ptbx-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1), 1–15.
- 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.
- Wang, Y.-C., Chen, T.-C. T., & Chiu, M.-C. (2023). An improved explainable artificial intelligence tool in healthcare for hospital recommendation. *Healthcare Analytics*, 3, 100147. <https://doi.org/https://doi.org/10.1016/j.health.2023.100147>
- Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review*, 115(1), 1–50. <https://doi.org/10.1215/00318108-115-1-1>



Interview Protocol

Introduction:

Interactivity to increase interpretability

- Utilization of interactive elements in an XUI to increase the interpretability of an XAI
- How does this work when applied to the case of FOKUS and their ECG case?

Interview Approach:

The interviewer starts with questions to gather background information about the participant. Afterwards the participant will have time to interact with the UI. Following this, open/reflective question will be asked where the interviewer may then probe with more specific questions. The extent of probing depends on the details given by the interviewee in open/reflective questions.

Goal:

- To analyse if the user is able to understand the elements present in the UI.
- To analyse if the interactive elements were helpful in increasing the interpretability of the XAI outputs.
- To analyse the extent to which the UI elements provide useful information for determining MI.
- To gather feedback on the UI

Questions:

1. Could you give an overview of your role and how ECG data relates to you [open]?
2. How would you classify your knowledge on ECG's [open]?
 - Would you be able to determine an MI?
3. Do you have any affinity with AI/XAI[open]?
 - Have you heard of XAI before?

TIME FOR INTERACTION 15MINS Short intro UI.

You will see 1 case of a patient in which an AI thinks the patient has experienced an MI. There will be 3 different XAI generated outputs which you can select from which will display their explanation for why the AI has marked it as MI.

During this time feel free to have a look around the UI and please see this as a think-aloud session. So whatever thoughts pop into your head, feel free to speak them out loud. If you don't feel like examining something thoroughly you can skip, but please have a look at everything.

1. Do you understand why the XAI labelled the cases MI? [probe]

-
- Yes -> go to question 2
 - No -> go to question 3
2. Did any of the elements in the UI help ? [probe]
 - Which elements were the most helpful in this case
 - Did you have a preference for which element was the most useful?
 - Were there any elements that were not useful?
 3. Can you explain why you didn't understand? [probe]
 - What would have helped you better understand?
 - Was there anything that was useful but just lacked more context/info?
 4. Did you find the added text information useful? [open]
 - Did it make the visual images more understandable? (added info or clutter) [probe]
 - Did you prefer the one over the other?
 5. Was it helpful to have the information presented in different bits? [probe]
 - Would have like a more deeper/complex explanation or would you even want a simpler one?
 - Was it useful to you that you could open and close certain elements in the UI?
 - Would you rather have static elements?
 6. Did you find it useful that you had multiple options to choose from?
 - Did you feel overwhelmed with the options? [open]
 - Would a counterfactual example would have helped in this case?
 7. (Do you think) Based off of the information provided in the UI, would you be able to determine an MI? (Depending on their expertise) [open]
 - Would you have enough information to do this under 5-6 minutes?
 - Would the added time pressure change your opinions that you've had so far?
 8. Is there any other question that you think I should have asked you or do you have any other feedback ?

B

Code Analysis

B.1. Code book

- Bias/prior
- Bigger image
- Blurry visuals
- Context
- Contrasting outputs by the XAI
- Counterfactual
- Decision under uncertainty regardless of outcome
- Deeper explanations
- Deeper underlying issues with AI
- Difficulty because of ECG
- Difficulty understanding XAI
- Disagreement in the diagnosis by AI
- Dislike for explanation 2
- Dislike for explanation 3
- ECG data used is confusing
- ECG knowledge
- Engineer perspective
- Experience
- Explanation selection
- Guiding on where to look
- Lacking ECG knowledge for MI diagnosis
- Layout
- Multiple explanations
- Multiple ways of explanation can be confusing if the explanations don't match up
- No diagnosis based on unreliable output
- Poor visualisation
- Preference
- Preference for colour instead of numbers

- Preference for explanation 1
- Preference for explanation 2
- Preference for explanation 3
- Progressive disclosure
- Simplicity
- Standard ECG visualisation
- Text
- The XAI output does not make sense
- The XAI/AI is focussed on the wrong things
- Too much information
- Underlying trust
- Understanding of output
- Unnecessary information
- Usage
- Use of different colours
- Useful for first time use
- Visualisation

B.2. Theme description and grouping

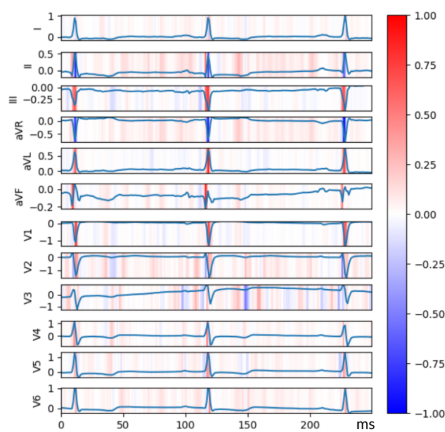
Table B.1: Themes identified through theme analysis, the corresponding codes and description

Theme	Codes	Code Names	Description
Complementary Naturalness	12	<ul style="list-style-type: none"> - Bigger image - Blurry visuals - Deeper explanations - Multiple explanations - Poor visualisation - Preference for colour instead of numbers - Simplicity - Text - Too much information - Unnecessary information - Use of different colours - Visualisation 	Combines implicit visual explanations and natural language rationales to make AI's inner workings more accessible to non-experts. This approach enhances user understanding and communication effectiveness by combining the precision of visual cues with the clarity and reassurance of textual explanations.
Responsiveness Through Progressive Disclosure	9	<ul style="list-style-type: none"> - Background information - Deeper explanations - Guiding on where to look - Multiple explanations - Progressive disclosure - Simplicity - Text - Too much information - Unnecessary information 	Use of hierarchical or iterative functions to offer explanations in stages, catering to the user's initial understanding. Aiming to balance information delivery by adjusting the level of detail to the user's comprehension, ensuring explanations are informative without being overwhelming.

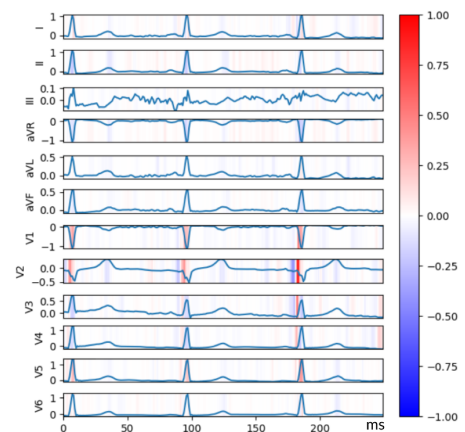
Flexibility Through Multiple Ways To Explain	11	<ul style="list-style-type: none"> - Contrasting outputs by the XAI - Counterfactual - Multiple explanations - Multiple ways of explanation can be confusing if the explanations don't match up - Preference - Preference for colour instead of numbers - Text - Too much information - Unnecessary information - Use of different colours - Visualisation 	People understand concepts differently, highlighting the absence of a one-size-fits-all explanation method. By employing a mix of visual, textual, and interactive explanations, the system can accommodate various learning styles, offering a more nuanced and adaptable understanding.
Sensitivity to context and mind	14	<ul style="list-style-type: none"> - Bias/prior - Context - Deeper explanations - Dislike for explanation 2 - Dislike for explanation 3 - ECG knowledge - Engineer perspective - Experience - Explanation selection - Preference - Preference for explanation 1 - preference for explanation 2 - Preference for explanation 3 - Underlying trust 	Tailoring explanations to the unique needs of each user, acknowledging that these needs evolve as understanding and trust develop. This personalised approach, which considers the user's preferences, prior beliefs, and cognitive processes, ensures explanations are both meaningful and effective, enhancing the user's engagement with machine learning systems.
Problems with XAI	11	<ul style="list-style-type: none"> - Blurry visuals - Context - Deeper underlying issues with AI - Difficulty because of ECG - Difficulty understanding XAI - Disagreement in the diagnosis by AI - ECG data used is confusing - No diagnosis based on unreliable output - Standard ECG visualisation - The XAI output does not make sense - The XAI/AI is focussed on the wrong things 	The challenges stemming from the incorrect application of ECG data in the XAI, which causes confusion and diminishes the XAI system's ability to effectively explain and support decision-making processes.
Environmental setting of the XUI	4	<ul style="list-style-type: none"> - Context - Ethical AI - Usage - Useful for first time use 	Examines the context, usage, and first-time utility of the XUI, addressing the topics of who uses it, when, where, and how. Highlighting the importance of adapting the XUI to specific situational needs and conditions.
Diagnosis	3	<ul style="list-style-type: none"> - Decision under uncertainty regardless of outcome - Lacking ECG knowledge for MI diagnosis - Understanding of output 	The ability of the participants to understand the outcome of the XAI and to make a diagnosis based on the insights
XUI design	2	<ul style="list-style-type: none"> - Guiding on where to look - Layout 	Preferences and feedback on the UI design part.

C

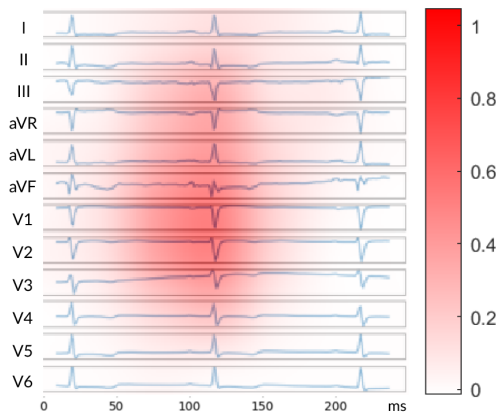
Overview XAI outputs



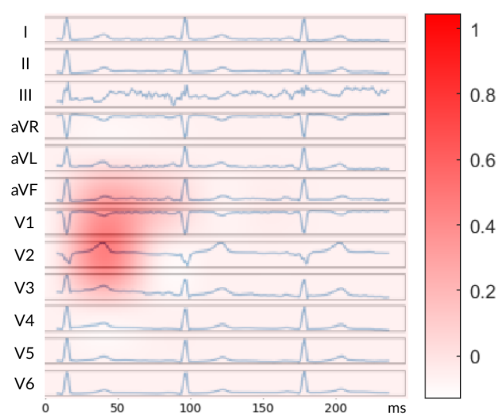
(a) SHAP



(b) SHAP



(c) Grad-CAM



(d) Grad-CAM

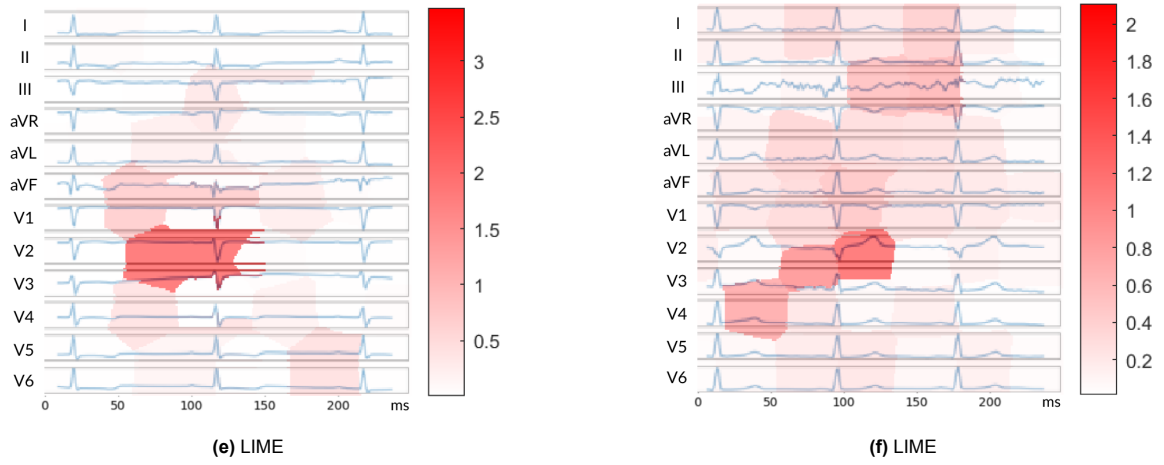


Figure C.1: Side by side view of the XAI outputs for two patients

D

XUI pages

Explanation 1 Explanation 2 Explanation 3 **MI DETECTED**

Patient information	
Patient ID	7591
Age	50
Sex	Female
Height (cm)	175
Weight (kg)	83
Date/Time	26/04 09:10

Prediction score ▾

On a scale from 0 to 1, where 0 represents the lowest level of certainty and 1 signifies the highest, the prediction score for this specific prediction is rated as a 1.

The model exhibits **maximum confidence** in its prediction and has the utmost certainty in the accuracy of the outcome.

XAI method ▾

The XAI method used here is SHAP (SHapley Additive exPlanations) which provides explanations for complex models using time series data. When applied to time series data, SHAP uses a unique approach to attribute the contribution of each time point or feature to the model's prediction. It considers all possible combinations of time points or features and calculates the Shapley values, which quantify the importance of each element based on their impact on the model's output. By utilizing the Shapley values, SHAP generates explanations that highlight the specific time points or features that significantly influenced the model's decision.

XAI output ▾

On the image on the left, certain points or areas have been marked with different colours. These colours represent the significance of those points in influencing the predictions made by the AI.

The points are categorized as follows:

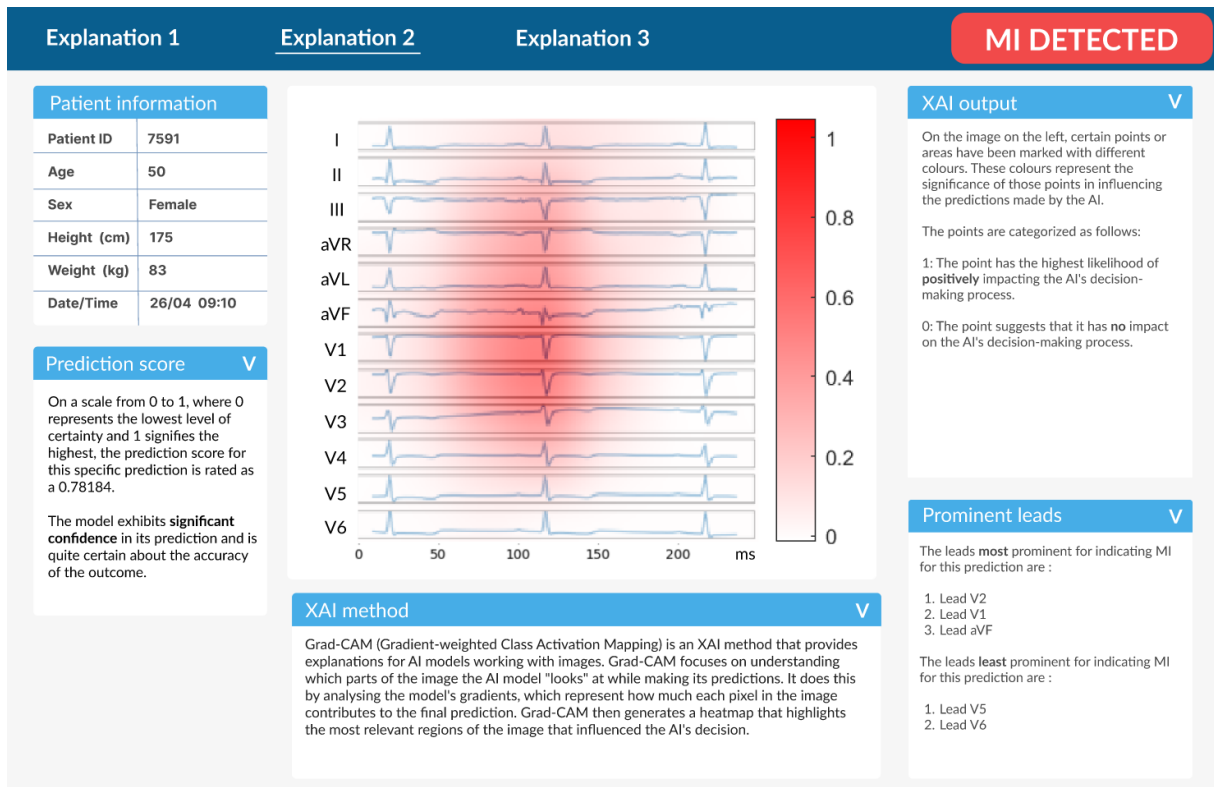
- 1: The point has the highest likelihood of **positively** impacting the AI's decision-making process.
- 0: The point suggests that it has **no impact** on the AI's decision-making process.
- 1: The point has the highest likelihood of **negatively** impacting the AI's decision.

Prominent leads ▾

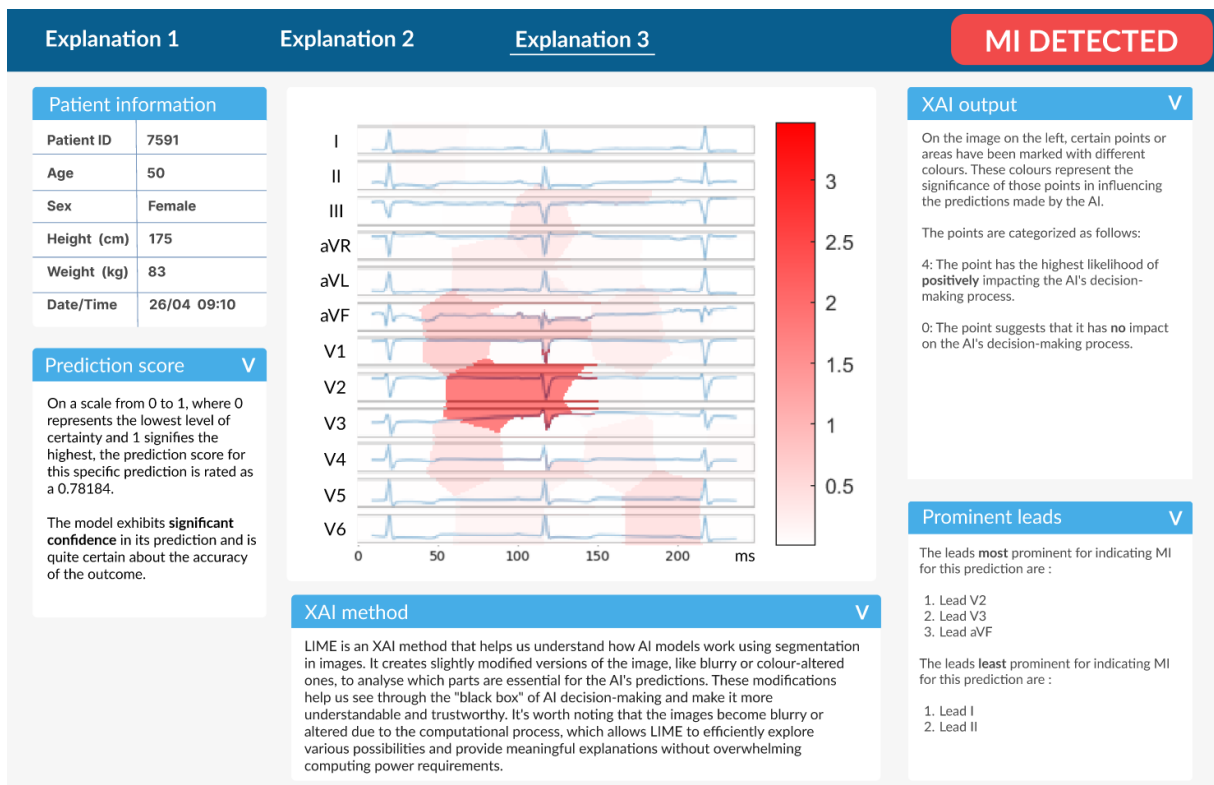
The leads **most prominent** for indicating MI for this prediction are :

1. Lead III
2. Lead II
3. Lead V2

(a) Patient 1, Explanation 1

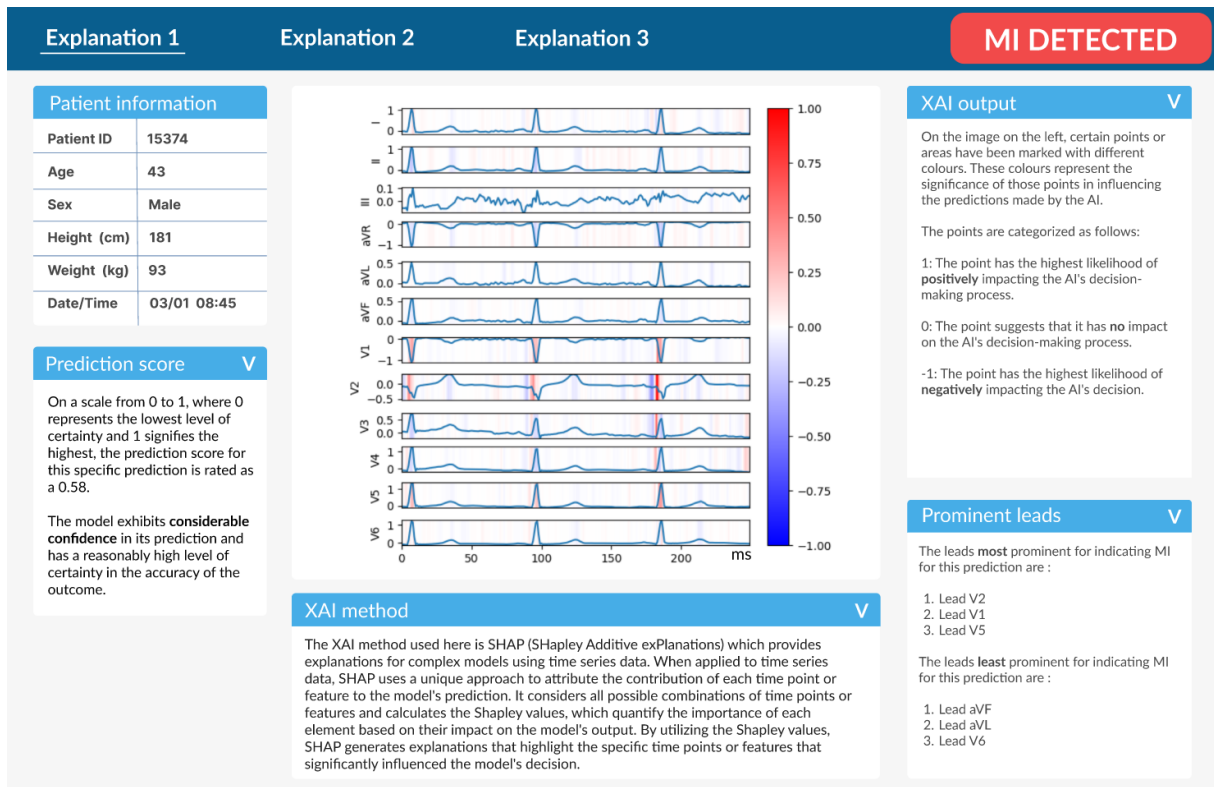


(b) Patient 1, Explanation 2

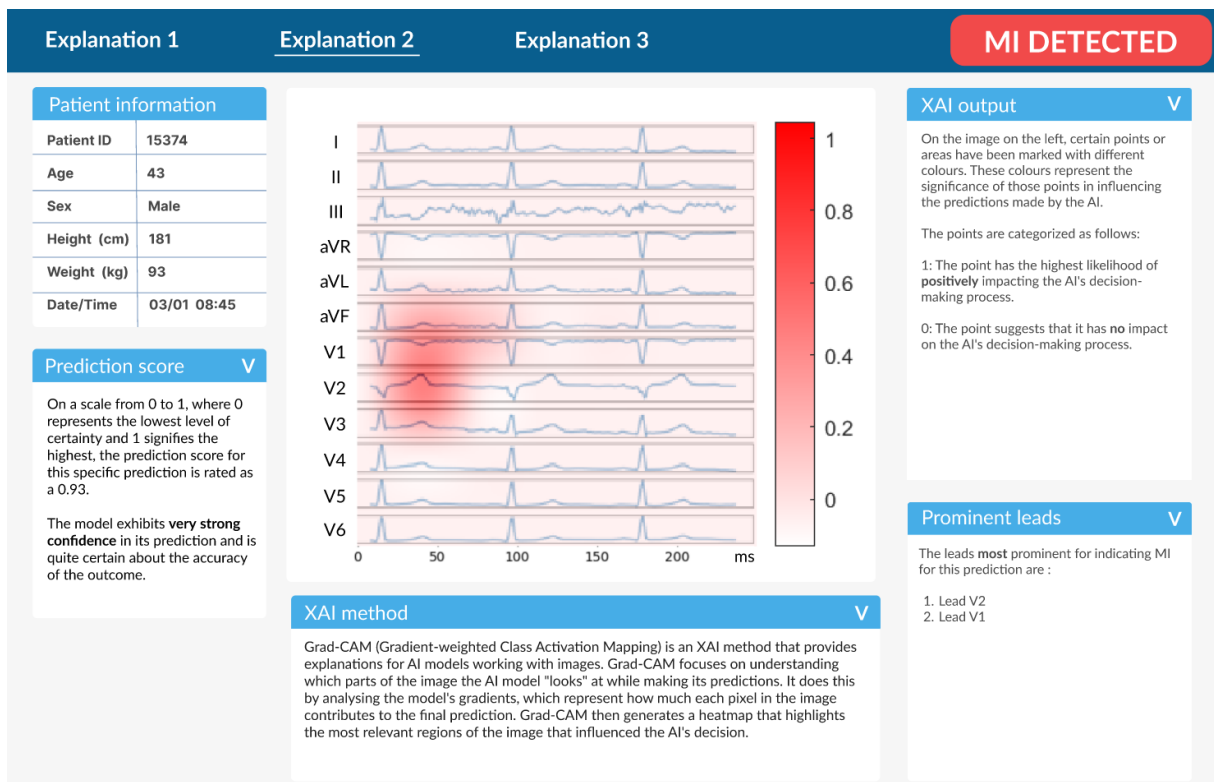


(c) Patient 1, Explanation 3

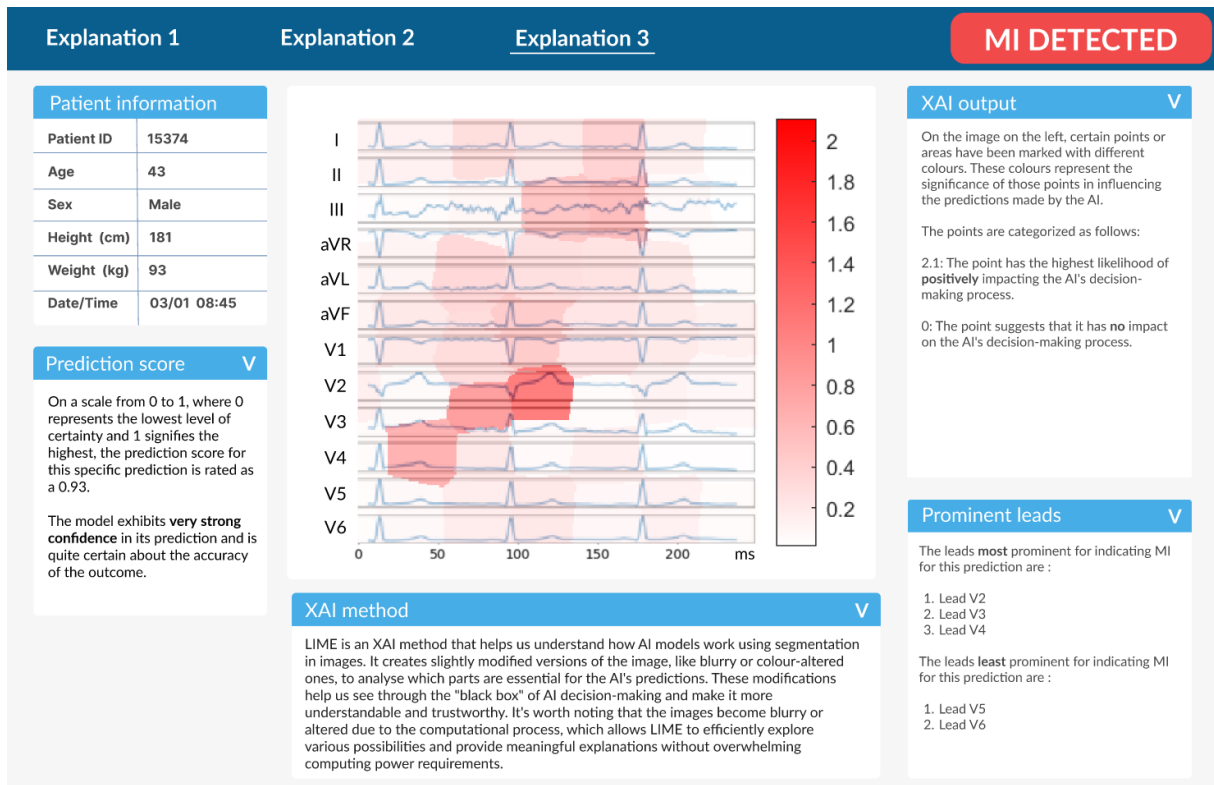
Figure D.1: XUI pages for patient 1



(a) Patient 2, Explanation 1



(b) Patient 2, Explanation 2



(c) Patient 2, Explanation 3

Figure D.2: XUI pages for patient 2