Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

Elise Almestad

Exploring Explainable AI Adoption in Medical Diagnosis and the Empowering Potential of Collaboration

Masteroppgave i Datateknologi
Veileder: John Krogstie
Juni 2022

NTNU
Kunnskap for en bedre verden

Elise Almestad

Exploring Explainable AI Adoption in Medical Diagnosis and the Empowering Potential of Collaboration

Masteroppgave i Datateknologi
Veileder: John Krogstie
Juni 2022

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

NTNU
Kunnskap for en bedre verden

NTNU
Kunnskap for en bedre verden

Department of Information Technology

TDT4900 - Master Thesis

Exploring Explainable AI Adoption in Medical Diagnosis and the Empowering Potential of Collaboration

Author:

Elise Almestad

Date: 22.06.2023

Abstract

The healthcare sector needs to include prominent examples of technology that seamlessly enables the successful integration of AI into clinical practice. While many systems designed for clinicians may function correctly, they often need to align with the clinicians' daily workflow, leading to limited adoption compared to what is intended. Also, because society is increasingly relying on AI systems to solve complex issues, at least in light of our recent experience with the pandemic in 2020, clinicians must have trust in these systems. Ethical dilemmas are created since the decisions made by clinicians impact many human lives, and there is a lack of trust and transparency along with an increase in the use of AI systems. The clinician must trust the AI system in order for the patient to trust that the doctor will make the appropriate choice if the AI system directs the clinician. Medical diagnosis is the the entry point in to the healthcare sector; therefore, many resources can be minimized by making a more accurate diagnosis or eliminating healthy patients.

The multicase study methodology in my master thesis facilitates data collection through firsthand observation of events and interviews with key stakeholders, including clinicians and developers. This approach facilitates a thorough examination of the challenges encountered by clinicians and developers of Explainable Artificial Intelligence (XAI) systems, offering multiple perspectives for comprehensive analysis.

Clinicians acknowledge that AI systems used for medical diagnosis should be transparent and explainable so that they can trust the results and understand how the system works. They are concerned about AI model biases such as failing to incorporate distinct demographics or datasets having uneven data quality. The study concluded that it is critical to solve these bias challenges and make AI outputs explainable and transparent to clinicians. Although the study did not completely analyze the specific criteria of the European Union Medical Devices Regulation (EU MDR), collaboration between developers and clinicians is necessary to achieve transparent Artificial Intelligence (AI) systems to the needs of the users.

Overall, this study contributes to understanding the challenges and potential strategies to mitigate them in the development and implementation of AI systems for medical diagnosis. By meeting these challenges, the field can make significant progress toward effectively leveraging AI technologies in healthcare and improving patient outcomes.

Sammendrag

Helsesektoren trenger å inkludere fremtredende eksempler på teknologi som sømløst muliggjør vellykket integrasjon av AI i klinisk praksis. Mens mange systemer utformet for klinikere kan fungere riktig, må de ofte tilpasses klinikerens daglige arbeidsflyt, noe som begrenser adopsjonen sammenlignet med hva som er tiltenkt. Ettersom samfunnet i økende grad er avhengig av AI-systemer for å løse komplekse problemer, spesielt med tanke på vår nylige erfaring med pandemien i 2020, må klinikere ha tillit til disse systemene. Etiske dilemmaer oppstår siden beslutningene tatt av klinikere påvirker mange menneskeliv, og det er en mangel på tillit og åpenhet sammen med økt bruk av AI-systemer. Klinikeren må ha tillit til AI-systemet for at pasienten skal ha tillit til at legen tar riktig valg hvis AI-systemet veileder klinikeren. Medisinsk diagnose er inngangsporten til helsesektoren; derfor kan mange ressurser minimeres ved å stille en mer nøyaktig diagnose eller eliminere friske pasienter.

Multicase-studiemetodologien i min masteroppgave fasiliterer datainnsamlingen gjennom førstehåndsobservasjon av hendelser og intervjuer med sentrale interessenter, inkludert klinikere og utviklere. Denne tilnærmingen muliggjør en grundig undersøkelse av utfordringene som klinikere og utviklere av forklarbar kunstig intelligens (XAI)-systemer står overfor, og gir flere perspektiver for omfattende analyse.

Klinikere erkjenner at AI-systemer som brukes til medisinsk diagnose, bør være transparente og forklarlige, slik at de kan ha tillit til resultatene og forstå hvordan systemet fungerer. De er bekymret for AI-modellers skjevheter, for eksempel manglende inkludering av ulike demografier eller datasett med ujevn datakvalitet. Studiet konkluderte med at det er avgjørende å løse disse skjevhetsutfordringene og gjøre AI-resultater forklarlige og transparente for klinikere. Selv om studiet ikke fullstendig analyserte de spesifikke kriteriene i den europeiske reguleringen for medisinsk utstyr (EU MDR), er samarbeid mellom utviklere og klinikere nødvendig for å oppnå gjennomsiktighet kunstig intelligens (AI)-systemer til brukernes behov.

Generelt sett bidrar dette studiet til å forstå utfordringene og potensielle strategier for å håndtere dem i utviklingen og implementeringen av AI-systemer for medisinsk diagnose. Ved å møte disse utfordringene kan feltet gjøre betydelige fremskritt i å utnytte AI-teknologi på en effektiv måte i helsetjenesten og forbedre pasientresultatene.

Preface

The Department of Computer Science at the Norwegian University of Science and Technology (NTNU) supported me in preparing this master's thesis. The master thesis examines an in-depth analysis through interviews with St. Olavs Hospital clinicians. These clinicians are engaged in studies to incorporate XAI into the ongoing development of future healthcare systems. Interviews with the developers participating in these projects were conducted to acquire a thorough insight, allowing for an examination of the interdisciplinary collaborative dynamics among project members. Finally, the study's findings aim to inform successful XAI integration strategies in future healthcare systems, enabling improved collaboration and aligning stakeholders' expectations.

I want to express my gratitude to my supervisor, John Krogstie, who guided me throughout my project assignment and my master's thesis. John has generously shared his expertise with me through conversations where I felt comfortable asking questions, discussing challenges and solutions, and providing me with the necessary and relevant information for my project assignment and master's thesis. Furthermore, he has facilitated my access to relevant material, such as articles and theory books, and contacts related to the three projects I investigated. The latter has been particularly crucial since only a few projects out there possess the specific characteristics I focused on. By setting up weekly meetings and internal deadlines for partial assignments, he has created a structured working environment for me, which has been valuable for a student writing a master's thesis alone. Our collaboration humbles me, and I will gladly recommend John Krogstie to future master's students.

I also want to thank all the participants who contributed to my case study. Additionally, I want to thank my classmates with whom I shared the master's room for an entire year. We have experienced excellent and challenging days but strived to uplift and motivate one another.

Table of Contents

List of Figures

List of Tables

1 Introduction

A number of medical applications of *Artificial intelligence* (AI) is reported. Many of these are based on *Machine Learning* (ML) and a challenge for taking these into use in clinical practice is the limited explainability of them. Thus *Explainable Artificial Intelligence* (XAI), Medical specialists should examine medical XAI applications. However, the majority of medical XAI apps lack XAI evaluation and medical expert review. Medical XAI apps should have well-designed *Human-Computer Interface* (HCI) and give medical specialists a plausible explanation for the provided results (Miró-Nicolau et al. 2022). As a result, substantial investigations are required to collect feedback from medical experts and examine possible opportunities and obstacles from their perspective. In this project, I will perform a study that will help to identify the gaps between the demands of XAI researchers and end-users in the real-world application of XAI systems (Payrovnaziri et al. 2020).

1.1 Motivation

In my master project, which was conducted last fall, the focus was on examining the challenges encountered in the relationship between medical professionals and the implementation of XAI in health systems for medical diagnosis. A comprehensive literature review revealed that various issues related to fairness, transparency, explainability, privacy, and transferability pose challenges in the interaction between XAI systems and medical professionals. The review also emphasized the significance of involving multiple disciplines and end users in the design process to effectively address challenges related to usability and system knowledge. Additionally, further research is required to establish precise definitions and standardized procedures for comparing and validating explanations. Active participation of medical professionals and collaboration among multidisciplinary teams are crucial for fostering advancements in XAI development.

The healthcare sector lacks prominent examples of technology that seamlessly enables the successful integration of AI into clinical practice. While many systems designed for clinicians may function correctly, they often fail to align with the clinicians' daily workflow, leading to limited adoption compared to what as intended. It is essential for a system to add value to the clinicians' workflow in order to effectively streamline processes and contribute to the overall value in clinical practice. Therefore, considering the perspective of clinicians is equally important as evaluating the actual performance of the system. To achieve successful integration of AI into clinical practice, involving clinicians in the validation of AI systems throughout the entire process is crucial.

The *European Union Medical Devices Regulation* (EU MDR) has introduced a new regulatory framework for medical devices in the *European Union* (EU), and is also applied in Norway. The aim is to align legislation with technological advancements and create uniformity across the EU. A significant change in the new regulatory framework is the requirement for more rigorous clinical evidence regarding safety, predictability and transparency (Huusko et al. 2023). After some scan-

dals within the medical device industry, the EU saw the need to strengthen the directives (Huusko et al. 2023). MDR emphasizes the importance of systematic clinical evaluations based on data to assess the benefit-risk profile of medical devices, and will under the evaluation evaluate bias, random error and lack of transparency in reporting and misinterpretation, is mentioned (Niemiec n.d.). The new regulations expand the scope, introduce stricter classification rules, and emphasize the need for more rigorous clinical evaluations before any medical device can be used in practice. Therefore is the EU MDR a important legalisation to gain knowloedge about, for both the developers creating medical devices and it is important for the developers to include the clinicians in the design-process.

1.2 Context

This master's thesis builds upon a project assignment conducted in the autumn of 2022, with a primary focus on exploring the challenges encountered by clinicians when utilizing XAI systems within the healthcare sector. The thesis was prepared with the support of the Department of Computer Science at the Norwegian University of Science and Technology (NTNU).

Employing a multi-case study methodology, this thesis delves into an in-depth investigation through interviews conducted with clinicians affiliated with St. Olavs Hospital. These clinicians actively participate in projects dedicated to the implementation of XAI within the ongoing development of future healthcare systems. To gain a comprehensive understanding, interviews were also conducted with the developers involved in these projects, thereby facilitating an exploration of the interplay and collaboration dynamics among team members. Additionally, this approach aimed to ascertain the extent to which the expectations of different roles harmonize with one another.

By pursuing this research, this thesis aims to contribute to the existing body of knowledge by clarifying the specific challenges faced by clinicians in the context of incorporating XAI into the healthcare domain. By examining all aspects of the subject matter, a thorough understanding can be obtained. Ultimately, the findings of this study aspire to inform successful integration strategies for XAI in future healthcare systems, fostering enhanced collaboration and alignment of stakeholders' expectations.

1.3 Research Question

For my master's thesis research, I chose three projects that are developing a system that strives to be both explainable and transparent, known as an explainable artificial system. As part of my research, I interviewed clinicians that will use the future prototype or have used the system in their current phase. I also interviewed one and two developers from each project to add a multidisciplinary perspective to my multicase study. Therefore, my research question will be:

Why or how do clinicians face challenges in the utilization of AI systems for medical diagnosis, and how will multidisciplinary cooperation contribute to mitigating these challenges?

1.4 Thesis contributions

The goal of this master's thesis is to shed light on the challenges that clinicians experience while utilizing XAI (Explainable Artificial Intelligence) systems, as well as the importance of their collaboration with the developer(s) involved in their development. The study provides qualitative data about the difficulties clinicians have when using XAI systems, as well as the nature of their collaboration with the system's developer(s).

1.5 Scope

The scope of this master thesis is focused on examining the contributions of clinicians at St. Olavs Hospital in Trondheim toward identifying challenges associated with the utilization of XAI systems in medical diagnosis. Specifically, the discussion will revolve around clinicians' role and collaboration with developers within their teams. The aim is to shed light on the experiences and perspectives of clinicians concerning these systems, highlighting their insights and potential areas for improvement.

1.6 Thesis structure

Table 1: Thesis Structure

Section Number	Section Name	Description
1	Introduction	This section aims to provide a comprehensive overview of the research topic by highlighting its relevance and underlying motivation. It serves the purpose of contextualizing the master thesis and introducing the research question. Furthermore, it outlines the scope within which the study will be conducted.
2	Background	The background section of this master thesis will review literature and theoretical concepts associated with the research question. It will include an exploration of fundamental aspects such as the hierarchy of AI and an introduction to The European Union Medical Device Regulation (EU MDR).
3	Method	In this section, I will explain how the case study was conducted, including the research methodology and strategy chosen. I will provide details on how the data was collected and evaluate the method used.
4	Multicase introduction	In this section, I will introduce the three projects/cases under investigation. I will provide a brief description of each project's goals and give an overview of the technology they utilize.
5	Result	This section will present the findings from the research questionnaire in a structured manner.
6	Discussion	The discussion section will provide an extensive review of the findings, and will be following the structure of the research questions.
7	Conclusion	The concluding section of this thesis will address the research question in light of the findings presented. The conclusion section will also discuss the limitations of the study and provide a reflection on them. Future research directions will be proposed.

2 Background

AI has grown significantly in importance in medical diagnosis in recent years, with good reason. It has helped lessen the pressure on a healthcare system threatened by a substantial influx of older patients by removing and reducing expenditures for routine duties and diagnosis and treatment costs. Accordingly, demand for workers can also be decreased, and maintaining sustainable development is more crucial than ever (Schwendicke et al. 2020). Although companies today need help with acquiring and preserving diverse data, they focus more on diversity. The health industry also has access to vast data sets that may be utilized to train machines (Alex 2022) (Schwendicke et al. 2020). The introduction of AI systems in the health industry has boosted efficiency, which has helped to increase the amount of time clinicians can devote to their patients. Therefore, the question of whether AI will transform and advance the healthcare industry has been replaced by the question of how this technology will be able to do so (Xue et al. n.d.).

Despite significant breakthroughs enabled by AI-driven technology in the healthcare field, the inherent transparency, commonly known as the "black box" problem, is a barrier to continued growth and innovation (Minh et al. 2021). The black-box problem affects transparent AI-driven systems, such as those that employ the DNN algorithm. The users of this algorithmic system, who are medical professionals, are unaware of what happens within the algorithmic process because it uses "hidden" and complex neural networks. As a result, healthcare professionals cannot assess the results offered, which breeds distrust in the systems (Arrieta et al. 2020). This is particularly problematic in the medical systems since choices can have far-reaching effects. Many diverse elements, including human ones like education level, attitudes toward technological advancement, and life experience, can all impact how clinicians perceive AI-driven systems. The system's characteristics, such as whether they are transparent and controlled, influence clinicians' trust when utilizing AI-driven systems.

Because society increasingly relies on AI systems to solve complex issues, at least in light of our recent experience with the pandemic in 2020, clinicians must have trust in these systems (Schwendicke et al. 2020). Ethical dilemmas are created since the decisions made by clinicians impact many human lives, and there is a lack of trust and transparency along with an increase in the use of AI systems. The clinician must trust the AI system in order for the patient to trust that the doctor will make the appropriate choice if the AI system directs the clinician (Alex 2022). Since medical diagnosis is the gateway to the beginning of a journey in the healthcare sector, many resources can be minimized by making a more accurate diagnosis or eliminating healthy patients (Y. Kumar et al. 2023).

This section will present the hierarchical framework of Artificial Intelligence (AI) to establish a contextual understanding of Explainable Artificial Intelligence (XAI). Furthermore, the inherent dilemma of Performance versus Explainability, which developers encounter during the development

of medical systems, will be introduced. Subsequently, section 2.2 will provide an overview of the fundamental aspects of The European Union Medical Device Regulation (EU MDR), followed by an examination of the challenges in relation to the EU MDR and XAI within the same section.

2.1 Hierarchy of AI

2.1.1 Machine Learning

ML is a subset of AI that enables machines to acquire knowledge from data without explicit programming (MD 2018). The main goal of ML is to train machines using available data and algorithms, allowing them to process information and make informed decisions (Jakhar and Kaur 2020). These ML algorithms exhibit dynamic behavior, as they can adapt and modify themselves when exposed to additional data. The fundamental aim of ML algorithms is to minimize errors and optimize prediction accuracy (Jakhar and Kaur 2020). The classical machine learning methods use input that experts within the field have specifically selected from the given data (Petry et al. 2020). That is, the subject experts must, with their prior knowledge, extract important features in advance so that they can be used as input in the machine learning algorithm. After processing the data, the algorithm generates an output(Petry et al. 2020). The classical machine learning methods usually only have one or a few layers between the input and output; they can also manage small-scale data sets with few features. For example, a linear regression algorithm calculates coefficients for each feature and thus gives insight into how much each feature is weighted in the output. This makes it easier to achieve explainability and transparency with such methods. See Figure 1 below.

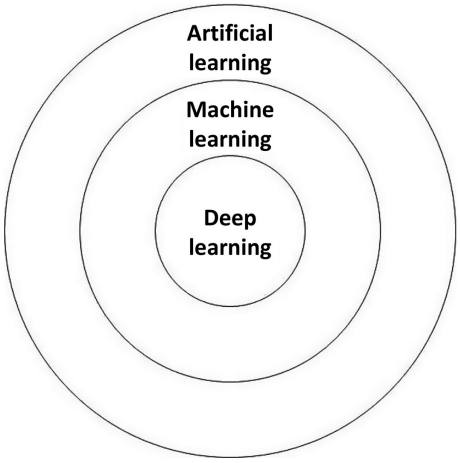


Figure 1: Hierarchy of AI, Machine Learning, and Deep Learning

2.1.2 Deep Learning

Deep learning (DL) is a specialized branch of machine learning that uses artificial neural networks. These networks have multiple hidden layers, allowing them to learn and represent data at several levels. (Jakhar and Kaur 2020). Unlike classical machine learning algorithms, deep learning models have a hierarchical structure with several interconnected layers between the input and output. This architecture enables the algorithm to perform complex calculations by gradually processing information through multiple layers. By utilizing this depth, DL algorithms can extract intricate features and gain a deeper understanding of complex patterns in the data. This hierarchical approach enhances the learning capacity and predictive capabilities of DL models, making them particularly effective in handling complex tasks and large-scale data sets (Jakhar and Kaur 2020).DL has been used in clinical imaging to analyze various types of medical images (Miotto et al. 2018). For example, DL models have been applied to brain MRI scans to predict Alzheimer's disease and its variations.

2.1.3 Performance VS Explainability

ML, particularly DL, has garnered significant attention in academic research. It has found applications in various domains, including medical imaging, offering ongoing advancements and inherent performance benefits (Miotto et al. 2018). Traditional ML methods and deep learning techniques have been employed to tackle complex tasks such as diagnosis and image classification, potentially enhancing the quality and efficiency of healthcare practices (Jakhar and Kaur 2020).

However, DL models have often been regarded as black boxes, posing challenges in generating explanations for their outputs within the field of XAI (V. S. Kumar and Boulanger 2021). The increasing complexity of deep learning approaches, coupled with the demands for transparency and explainability, has led to a necessity for more significant agreement on implementing XAI in the healthcare field (Jakhar and Kaur 2020).

XAI is an approach to AI that seeks to address this problem by making AI models more transparent and interpretable. XAI aims to enable humans to understand how AI models work, what factors contribute to their decisions, and how they can be improved or made more reliable. Using XAI techniques, researchers can build models that achieve high accuracy and provide explanations for their decisions. This makes it easier for humans to trust AI models and to use them in real-world applications, such as medical diagnosis (Jakhar and Kaur 2020).

Within the healthcare domain, AI-based systems must accommodate the perspectives of medical practitioners in order to be considered explainable. Designing domain-agnostic systems with XAI that can cater to multiple perspectives is a complex task due to the contextual nature of explanations. Stakeholders may prioritize different aspects of the system's operation and require tailored explanations based on their needs (Jakhar and Kaur 2020).

Furthermore, interpretability and explainability extend beyond the operation of ML models and encompass the entire workflow used to train these models. This encompasses data preprocessing steps, the selection of ML models, and evaluation criteria. The workflow provides technical insights to ML engineers while offering medical practitioners a deeper understanding of the underlying data, model interpretation, and relevant performance metrics for medical diagnostics (Jakhar and Kaur 2020).

Addressing the challenges associated with DL and explainability necessitates the development of methods within XAI that enable the generation of explanations and insights into the decision-making processes of models. This is crucial in establishing trust and acceptance of AI systems within clinical practices. Further exploration and advancement of XAI methods are required to make DL and other ML techniques more explainable and transparent. This will foster trust in AI systems by medical professionals, ultimately strengthening the application of AI in the healthcare sector (Jakhar and Kaur 2020) (Bharati and Mondal 2022).

2.2 The European Union Medical Device Regulation

Interpretability and explainability link clinicians' and doctors' domain expertise with specific results. Moreover, explainability aims at functional benefits and adds to clinical confidence and consistent compliance with legal and ethical requirements. These refer to regulations such as the European Union's General Data Protection Regulation, which enforces the right of patients to receive transparent information about a decision's origin, or the European AI Act, which introduces a regulatory and legal framework for AI systems. Lastly, identifying errors, limitations, and potential biases is essential in developing and applying AI systems and must be addressed accordingly Borys et al. 2023. For that reason, after the enthusiasm to achieve high performance, the demand for interpretability and explainability of AI has experienced a significant resurgence over recent years, promoting the formation of a new research field known as XAI.

Regulators have expressed concerns about medical AI devices and networks in the healthcare industry (He et al. 2019). To increase patient safety associated with medical devices, the EU introduced the new Medical Device Regulations (MDR) in 2017, which was applied in May 2021, replacing the previous Medical Device Directive (MDD). The new regulation considers the technological advancements that medical equipment is undergoing and seeks to settle related issues (EU n.d.(a)) (EU n.d.(b)). The MDR classifies medical devices according to the risk levels it poses to society and the individual user. Medical device rules must be more stringent the more significant the risk the device poses. The steps for clinical evaluation of medical devices are outlined in the MDR and are further explained in MEDDEV 2.7 (Wilkinson and Boxtel 2020).

The new improvements have also increased the emphasis on Medical Device Software (MDSW) and classification standards that employ a risk-based approach, prioritizing user security, particularly

for software (Beckers et al. 2021). These modifications necessitate re-certification of medical equipment and stricter development and maintenance practices (Bianchini and Mayer 2022). The patient safety aspect of the rule encompasses various components, such as ensuring data transparency, implementing follow-up monitoring procedures, and facilitating the accessibility of implant-related information. The strategy for implanted information has yet to be developed; hence there are no unique guidelines for data requirements (Vasiljeva et al. 2020).

The MDR also expressly requires the maker to implement a risk management strategy for the whole life cycle of a device. This entails lowering the device's risks, incorporating user feedback, and balancing risks with the technology's positive aspects. In this approach, the involvement of innovators, clinicians, and researchers in the design phase is essential. It also emphasizes the importance of collaboration and participation of all stakeholders, including the scientific community, in implementing the MDR's criteria. This emphasizes the involvement of researchers and clinicians in the innovation process as inventors, developers, and validators. The MDR will improve the safety and performance of medical AI devices on the European market, but only if countries implementing the EU MDR enforce it properly (Niemiec n.d.). Furthermore, stakeholders must create awareness and foster conversations about the EU MDR.

> *"Medical devices have a fundamental role in saving lives by providing innovative healthcare solutions for the diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease" - EU (EU n.d.(b))*

> *"Ensure a robust, transparent and sustainable regulatory framework" - EU (EU n.d.(a))*

Because medical equipment must comply with the EU's MDR, the identified challenges to explainable artificial intelligence will be linked to this legislation; Evaluating the medical device should consider any potential confounding influences, bias, random error, lack of reporting transparency, and misinterpretation (Niemiec n.d.). Since the EU MDR aims to increase patient safety, handling the patient's data will also be an aspect that will be investigated when identifying challenges in this master thesis.

2.2.1 Bias

A potential source of mistrust in the systems lies in the fact that there is a likelihood of the system being influenced by bias. (Campos Aranovich and Matulionyte 2022). Algorithms may reinforce or prolong preexisting biases in the data they are trained on. For instance, an AI system may get biased results against a specific group of people if it is trained on data skewed against that group. Aranovich says this may have adverse consequences, such as discrimination or unequal treatment. Projects that can eliminate data biases must be supported to address this issue. AI systems must be trained on a variety of representative data sets. Furthermore, it is crucial to ensure fairness is

considered while developing and evaluating AI systems and that sufficient measures are in place to monitor and rectify any fairness issues that may arise (Campos Aranovich and Matulionyte 2022). There are various types of bias related to AI systems, each with its own set of problems. Physical bias occurs when the design of a medical device favors certain demographic groups based on physical characteristics such as skin color (Martinho et al. 2021). Interpretation bias happens when medical equipment draws incorrect inferences based on readings, while computational bias happens when training data sets do not adequately reflect the people in society (Martinho et al. 2021). The risk of bias is so substantial, and for the explanation to be helpful in a medical setting, healthcare providers must inform patients of this risk of bias and any potential implications this can give in their medical diagnosis process (Martinho et al. 2021). The future development of explainable AI needs to focus on this to ensure that the new XAI systems are conscious of bias. Since bias cannot be eliminated, it is crucial to look at the limitations bias will place on XAI systems in order for the systems to function effectively (Martinho et al. 2021).

2.2.2 Transparency

One disadvantage of employing non-transparent systems and procedures is that you risk undermining the patient's autonomy, which weakens the doctor-patient relationship (Amann et al. 2020). Doctors, like patients, require knowledge regarding not just the presented outcomes but also the assumptions and causalities underpinning these results (Amann et al. 2020). If this does not proceed, the patient's confidence may be jeopardized since a "black box" cannot be trusted, and one's willingness to follow the clinical suggestions may decrease as a result (Amann et al. 2020). For example, suppose a patient finds out that their medical condition has been influenced and chosen by a non-transparent system. In that case, the patient may demand an explanation of the result that has been given. If the system is non-transparent, medical professionals cannot explain it. As a result, a non-transparent system will be challenging to use when a patient requests extensive information and may jeopardize future voluntary participation from patients. To avoid jeopardizing future voluntary participation, it is critical to follow ethical explainability requirements (Amann et al. 2020). Ethical explainability requirements are created to ensure that the explanation will be communicated optimally, and then the AI system becomes opaque so that the patient can trust it. This implies the patient believes the system's advice is the best option. In a doctor-patient interaction, it is an ethical need and prerequisite that systems that assist medical decision-making provide explainability.

A transparent artificial intelligence system can offer explanations in the form of visual images or textual explanations (Amann et al. 2020). To use such technology and provide patients with the most accurate diagnosis, medical professionals must be able to evaluate these explanations rather than unthinkingly believing the system. In addition, the medical professionals need to understand the system's functioning. Clinicians have the option of interpreting the data and determining

whether or not the results are reliable. Suboptimal explanations are called when the explanation is not unambiguous (Rasheed et al. 2022). Knowledge of the system enables clinicians to tailor their suggestions to all of the specific diagnostic conditions they will experience in practice with patients (Rasheed et al. 2022). This will assist in reducing the danger of misdiagnosis or false hope, but it will also help to exclude unnecessary action by employing common sense and professional judgment (Beil et al. 2019).

This raises another challenge: how a clinician should interact with the AI system if the doctor disagrees with the result (Grote and Berens 2020). It is argued that a non-transparent AI system lacks accepted epistemic authority since a black-box model cannot increase a doctor's ability to predict. However, it limits the clinician's skills (Grote and Berens 2020). Other reasons why it is not adequate to exclusively look at the system's results in clinical situations include that it might lead to "defensive medicine." Defensive medicine entails treating or other medical activities to avoid legal complications rather than performing treatment or other medical actions for the patient's benefit (*Defensiv medisin og hvordan det påvirker helsekostnader", url = "https://no.approby.com/defensiv-medisin/a meeting* n.d.). If defensive medicine is used in practice, it will undermine the doctor's autonomy and ethical foundation. Regarding epistemic authority, the best outcome for patients when the AI system and the doctor disagree will occur only when the medical professionals can explain the clinical evaluation (Grote and Berens 2020).

2.2.3 Explainability

One challenge with explainability is that it must attain both interpretability and completeness in order to be optimal (Rasheed et al. 2022) (Martinho et al. 2021). Often, a system can only provide a suboptimal explanation. It is common to utilize visualization approaches to explain an AI system's decision; however, the explanations provided by these methods are not always optimal for medical professionals. SHAP (Shapley Additive explanations) is an algorithm that employs game theory to generate values that inform us how the values are distributed in the prediction result (Chu n.d.). A survey was conducted to see if the SHAP explanations enhanced the doctors' decision-making skills. The findings revealed that the SHAP explanations did not improve the doctor's decision-making abilities (Weerts et al. 2019). Another algorithm, LIME, is an algorithm that generates a list of explanations for each feature value's contribution to the model prediction (Chu n.d.). The LIME explanations were also presented in a survey to compare the algorithm's explanation performance against ten medical professionals who provided their explanations (Weerts et al. 2019). According to the ten medical professionals, the observation resulted in the LIME algorithm exposing several irrelevant qualities to decision-making. Based on these findings, the above mentioned algorithms will not necessarily help decision-making (Weerts et al. 2019).

Even though the exact principles or extent of the application of explainable AI are unclear, fairness and explainability are two qualities that are generally acknowledged in the creation of artificial

intelligence in the healthcare industry (Campos Aranovich and Matulionyte 2022) (Martinho et al. 2021). A study was conducted in Portugal, the Netherlands, and the United States, where various medical professionals participated in evaluating statements regarding ethical artificial intelligence (Campos Aranovich and Matulionyte 2022). The study aimed to provide insight into ethical problems that should be considered while planning and enhancing breakthroughs in artificial intelligence in the health industry. According to the study's findings, participants stated that medical professionals should be included in the design process of AI systems to avoid problems such as suboptimal explanations (Campos Aranovich and Matulionyte 2022). Several studies have found that for technologies to be used to their full potential, clinical education should contain more information about using such systems (Martinho et al. 2021). The AI systems and the medical professionals share responsibility; the AI systems must provide the best explanations possible, but the medical professional who will use the technology must be as knowledgeable as possible about how to use the system because the machine will lack contextual knowledge and the ability to read social codes (Martinho et al. 2021). That is why doctors must engage in the design process, where they may offer advice to adapt and improve the explanation algorithms, allowing both parties to execute and carry out their job responsibilities to the best of their abilities.

An optimal explanation, in which clinicians participate in the design process, aids in creating trust in the system. As previously stated, the explanation should be complete and interpretable so that no suboptimal explanation misleads the end user. In addition, it should also provide information regarding the reason for any inaccuracy in the result to promote fairness and trust. A framework or standards for evaluating the correctness of an explanation have yet to be developed, making it risky to apply XAI in a medical environment when such measurements have yet to be implemented. The XAI technology's positive effects can then be transformed into negative implications.

Another challenge that XAI has is finding the right balance between algorithm accuracy and a valuable and optimal explanation (Rasheed et al. 2022). Traditional deep learning methods within machine learning are one of the reasons why achieving a transparent system with interpretable and detailed explanations is challenging (Rasheed et al. 2022). Many complex models in the healthcare sector produce accurate results, and the healthcare industry is dependent on these precise findings, but this goes beyond the model's explainability, creating a conflict between the algorithm's accuracy and the model's explainability (Rasheed et al. 2022). In such circumstances, an inherent and explanatory approach to the complicated and correct model may be a solution. However, this is a complex process (Rasheed et al. 2022).

2.2.4 Privacy

One of the most challenging aspects of privacy in XAI is ensuring that sensitive data is not revealed in the AI system's explanations. This is incredibly challenging when the AI system makes decisions based on input from many sources since verifying that all sensitive information is appropriately

safeguarded can take time and effort. Furthermore, the AI system's explanations may be utilized to reverse the underlying data or methods, potentially resulting in a breach of privacy.

Another privacy challenge in XAI is balancing the requirement for transparency and interpretation with the need to preserve sensitive data. In some circumstances, offering thorough explanations of an AI system's conclusions or forecasts may necessitate the exposure of sensitive information, putting privacy at risk (Arrieta et al. 2020).

Failing to secure patients' privacy may jeopardize the ethical orientation of artificial intelligence and the doctor-patient relationship (Martinho et al. 2021). Although adhering to the four pillars (Portal n.d.) of medical ethics is essential, it is now necessary to establish a framework and laws for accountability, data, and product certification. To address the issues associated with privacy, it is also suggested that AI systems in health be evaluated in randomized clinical trials (Martinho et al. 2021). A participant in a survey for clinicians in the Netherlands, Portugal, and the United States emphasized this point, saying that because the results of the AI system directly affect the patient's health, these systems should be treated in the same way as medicines, that is, to iterate through the same high standards as new medicine and new treatment strategies (Campos Aranovich and Matulionyte 2022).

Overall, preserving privacy in XAI necessitates careful consideration of the AI system's data and methods and the explanations supplied by the system. Additional safeguards or constraints may be required to protect sensitive information while allowing the AI system to deliver clear and interpretable explanations for its actions (Arrieta et al. 2020).

3 Research Method

In this section, I will explain how the case study was conducted, including the research methodology and strategy chosen. I will provide details on how the data was collected and evaluate the method used.

3.1 Research Method

XAI has gained significant prominence and continues to be a relevant topic in the field. This research concentrates on three ongoing projects that are actively developing distinct XAI systems, presenting a valuable opportunity to observe clinicians' utilization of these systems and conduct interviews with both clinicians and system developers. By investigating multiple cases, the study aims to furnish compelling evidence and ensure the robustness of the overall research findings (Yin 2017). In the context of a multiple-case study, two types of replications can be employed: theoretical replications that predict contrasting results and literal replications that predict similar results (Yin 2017). For this particular study, three literal replications have been chosen, enabling a comprehensive analysis of XAI implementation in medical contexts.

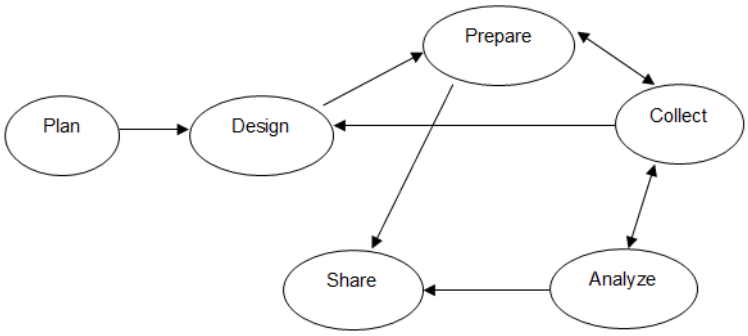


Figure 2: The Process of Case Study Method

The decision to adopt a multicase study approach in this master's thesis is underpinned by several factors. Firstly, the research attempts to explore recent developments in the application of XAI systems for medical diagnosis, with a specific focus on their functionality and associated challenges. Case studies are an excellent tool for diving deeply into these processes throughout time, going beyond the restrictions of simple frequency or occurrence analysis. (Yin 2017). Secondly, the case study methodology facilitates the collection of data through firsthand observation of events and interviews with key stakeholders, including clinicians and developers. This approach facilitates a thorough examination of the challenges encountered by both users and developers of XAI systems, thereby offering multiple perspectives for comprehensive analysis (Yin 2017). The process is iterative, as shown in 2.

3.2 Screening candidates

I selected three specific cases based on several factors. Firstly, these cases were chosen because they met the criteria for this master's thesis and were available through the contacts of my supervisor, John Krogstie, both in the medical field and at NTNU. To address my research question, it was essential to gain knowledge about the different systems, their utilization, and the collaboration between clinicians and developers within the chosen projects. Each case had a lead clinician who held primary responsibility and would be involved in testing the prototype in the future. In the Proviz project, I contacted the main clinician, a radiologist, and the two developers who have been engaged in the project since the beginning. For the DeepInMotion project, I reached out to the primary clinician, a physiotherapist, and the corresponding developer, who is currently pursuing a PhD. DeepInMotion is an extension of the original InMotion project, and thus the developer did not join the project from the beginning, unlike the clinician. Lastly, in the third project, I contacted a clinician who is a PhD student in medicine and the corresponding developer who was also involved. I directly approached all participants, and they voluntarily agreed to participate in in-depth interviews and demonstrate the use of the systems, enabling me to observe their functionality. Given that my investigation focuses on the multidisciplinary collaboration within these projects, it was crucial for me to interview both the clinician and the developer involved in each respective case.

3.3 Data collection

By employing interviews and incorporating multiple sources of evidence, this case study aims to provide a comprehensive and nuanced understanding of the research phenomenon within its specific context.

3.3.1 Interviews

In this case study research, interviews have been the primary source of data collection, specifically in the form of a semi-structured interview. Two interview guides were prepared before the interviews; In the respective order, the appendix includes an interview guide in English for developers (Appendix A) and one interview guide in English for clinicians (Appendix B). Interviews offer several strengths, including their targeted nature, allowing the researcher to focus directly on the topics relevant to the case study, and their ability to provide insightful information, including explanations, personal views, perceptions, attitudes, and meanings. However, interviews are not without their weaknesses, which include potential bias arising from poorly articulated questions, response bias, inaccuracies due to poor recall, and reflexivity, where interviewees may tailor their responses to align with the interviewer's expectations (Yin 2017). Furthermore, while the interview instructions are written in English, the interview participants were given the option of responding

in Norwegian if they wanted to do so. However, this opens the door to different interpretations and translations of the questions, which could lead to biased conclusions.

Two specific methods of data collection were employed in this case study. Firstly, a series of short case study interviews were conducted, involving approximately 5-6 participants. These interviews served the purpose of corroborating previously established findings while not addressing broader, open-ended topics (Yin 2017).

3.3.2 Observation

Secondly, the researcher observed the technological systems associated with the three cases. By directly observing these systems, a broader perspective was developed, albeit with the limitation that the systems were not fully matured or formal prototypes. This limitation arises from the challenges posed by the current landscape of XAI systems, where regulations in the EU may be vague or not yet implemented, resulting in slower deployment processes.

3.4 Analyzing the case study evidence

First i transcribed the interviews and imported them to the Nvivo, which is a computer-assisted tool. Specifically Nvivo, were utilized to analyze the collected data. Nvivo provides the researcher with the ability to effectively handle large volumes of textual data while employing coding skills and techniques (Yin 2017). A qualitative analysis was conducted, so the interviews conducted was transcribed, highlighted categories were identified and statements were coded (Oates et al. 2022).

The qualitative data for analysis was primarily derived from interviews. Within the Nvivo software, the researcher input the textual data and defined a set of codes, allowing the software to identify all relevant words and phrases that matched the specified codes. Through boolean searches, the software enabled the identification of multiple combinations of codes within the data files, a process that could be iteratively performed. Yin argues that utilizing the Nvivo database enhances the reliability and ensures the ongoing quality of the results (Yin 2017).

Once the coding process was complete, the process of studying the output started, focusing on patterns such as the frequency of codes or code combinations. This analysis serves as a foundational step, providing a basis for answering lower-level research questions pertaining to ”how” and ”why.” (Yin 2017).

To initiate the analytical strategy, the researcher conducted searches within the coded data, aiming to identify patterns, gain insights, and uncover relevant concepts. In summary the qualitative analysis involved transcribing the interviews conducted and identifying relevant categories, followed by coding the statements based on the method suggested by Oates (Oates et al. 2022). This top-down approach enables a systematic exploration of the data, facilitating the extraction of

meaningful findings in response to the research objectives.

3.5 Evaluation

This section aims to critically evaluate the reliability and validity of the research methodology utilized and the data collected for the current study.

3.5.1 Method Strategy and Limitations

The research design employed for this study entailed the selection of a multicase study approach, although an alternative research method such as ethnography could have been considered. Ethnographic research, characterized by its qualitative nature, entails an immersive and in-depth examination of cultures and social groups through extensive fieldwork (Oates et al. 2022). However, the adoption of an ethnographic approach would have necessitated a substantial allocation of time and resources, owing to the demanding nature of fieldwork and data collection procedures.

Triangulation, as employed in this case study, entails the utilization of multiple data sources or methods, specifically interviews and observations (Yin 2017). By incorporating these complementary approaches, a broader range of perspectives and insights can be obtained. Triangulation plays a fundamental role in reducing bias, enhancing data reliability, and facilitating a more comprehensive understanding of the phenomenon under investigation. Through the process of cross-checking, the validity and credibility of the study's conclusions are reinforced, thereby fortifying the overall robustness of the research outcomes (Yin 2017).

In this qualitative research methodology, the chosen approach involved conducting semi-structured interviews. This interview format offers the interviewer flexibility by providing a predetermined interview guide, as outlined in the Appendix, while also allowing for the inclusion of follow-up questions in areas of particular interest. This flexibility proved valuable in the context of investigating the three systems, each with distinct starting points. The semi-structured format enabled the interviewer to adapt their approach to each participant, ensuring a personalized and meaningful engagement.

However, it is important to acknowledge that the use of a semi-structured interview format introduces certain limitations. This format introduces variability in the data collected, as different researchers may employ unique styles and utilize diverse follow-up questions during the interviews. Additionally, the data obtained from participants heavily relies on their self-reporting, which can be influenced by recall bias or social desirability bias. These biases have the potential to impact the accuracy and reliability of the information provided by participants. Awareness of these limitations is crucial in interpreting and analyzing the data collected in this study.

Engaging in a qualitative multicase study as a first-time interviewer entails specific limitations

and challenges that should be acknowledged and addressed. Developing and refining the necessary skills for effective data collection is a continuous process that requires time and practice (Oates et al. 2022). Furthermore, it is crucial to be mindful of the potential introduction of bias, such as through leading questions or personal assumptions, which can influence participants' responses and undermine the objectivity and reliability of the collected data. By proactively addressing these limitations through thorough preparation and awareness, it is possible to mitigate the aforementioned challenges and enhance the quality and validity of the study's findings.

3.5.2 Ethical practice

When exploring a contemporary phenomenon in its real-world context, it is crucial for researchers to adhere to ethical practices similar to those followed in medical research. When conducting a case study, it is essential to exercise careful consideration and sensitivity in safeguarding the rights and well-being of participants. This entails a set of measures to be implemented:

Obtaining informed consent from all individuals involved in the case study is imperative. This involves providing detailed information about the nature of the study and seeking their voluntary participation.

Secondly, ensuring the protection of participants from any harm is of utmost importance. This includes refraining from the use of deception during the study, prioritizing their well-being throughout the research process.

Finally, upholding the privacy and confidentiality of participants is paramount. It is crucial to ensure that their involvement does not expose them to unwanted consequences, such as being included in future studies without their knowledge or consent, regardless of whether conducted by the researcher or others.

To reaffirm the commitment to these ethical principles, the research plan has obtained formal approval from SIKT/NSD, providing further assurance of adherence to these ethical guidelines.

4 Introduction to the Multicase

The three selected projects in this master thesis exhibit notable similarities that deserve emphasis. Each of these projects relates to the healthcare sector and strives to develop a tool that complies to the guidelines outlined in the EU MDR. As a result, the primary objective of all three projects is to create innovative tools capable of effectively reducing the time required for medical diagnosis within the healthcare system. Moreover, these tools aim to minimize the resource utilization associated with this crucial process.

This section of the thesis aims to provide a comprehensive overview of three distinct cases involving XAI systems that i will be investigating in thesis. The focus will be on examining the development of system technologies in relation to the key principles of explainability, transparency and bias. Specifically, the cases under examination are base on the development of the Proviz project, Deep in motion project, and Automated Assessment of Left Ventricular Function by Advanced Ultrasound project. Each of the cases will be described, shedding light on the strategies and technology employed to ensure the systems' explainability and transparency.

4.1 Proviz - Case A

The Proviz project aims to create a system that will help clinicians diagnose patients who are at risk for prostate cancer. Patients that who are classified as being in risk for prostate cancer undergo the most extensive treatment protocols in Norway. This decision support system therefore aims to improve the precision , and reduce the number of patients who have to undergo a prostate biopsy. The system will use AI techniques, such as deep learning algorithms, with models that will be trained on data sets consisting of data from more than 1,600 Norwegian patients from MRI scans. The Proviz project consists of both experts on the technical side; developers, and AI researchers, but also clinicians who are experts in this clinical field.

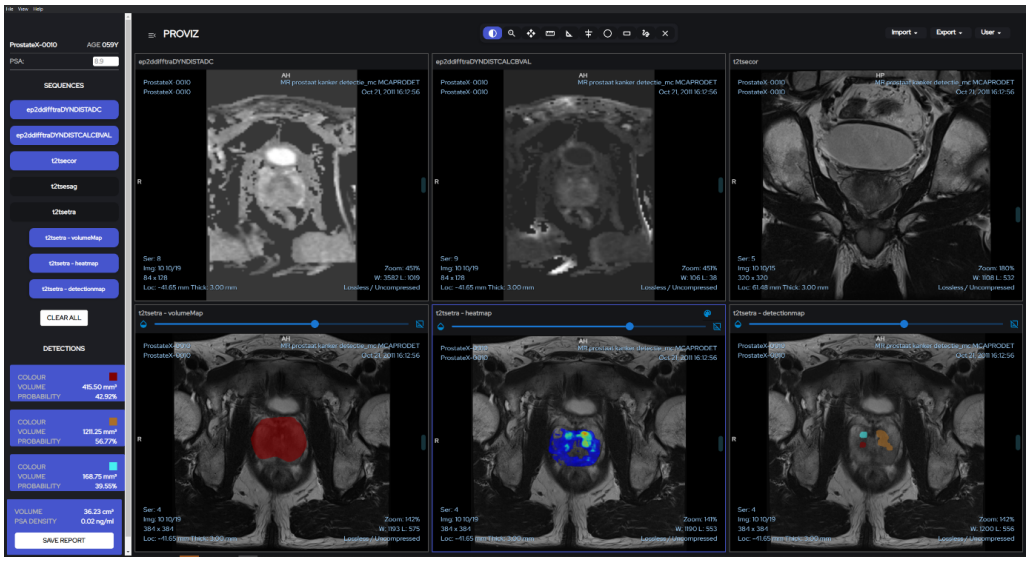


Figure 3: Several frames with pictures of the prostate. The colored areas indicate the degree of risk of cancer.

4.1.1 Overview of the technology

The XAI system uses a combination of deep learning algorithms and classical machine learning algorithms to harness the benefits and disadvantages of both types of algorithms within medical image analysis. The main goal is to segment the prostate and provide anatomical features achieved through the application of deep learning techniques. Deep learning is beneficial in accomplishing this because it is highly effective in accurately segmenting the MR-pictures of the prostate, but when it comes to making clinical decisions, such as determining whether a prostate is likely to have cancer or not, the system incorporates classical machine learning methods. This decision is made to ensure a level of explainability in the system's outputs.

The XAI system can be described as a combination of pre-modeling and explainable modeling stages. In the pre-modeling stage, the system focuses on extracting radiomic features from the data.

Radimoic features are features that allows for an extraction of clinically relevant information from radiologic imaging (Tomaszewski and Gillies 2021). By employing this approach, the system gains prior knowledge about the features and their meanings before proceeding to the modeling phase. By investigating the relation between the output results and the input features, the system aims to gain significant insights. For example, the system evaluates whether the feature values related with regions that indicate the presence of cancer are high or low. This study improves the transparency of the system's outputs. Additionally, the system's inherent explainability is attributed to its utilization of classical machine learning methods, which are based on linear regression modeling principles. Figure 4 presents the output of the system

4.2 DeepInMotion - Case B

Deep in motion project aims to provide secure and efficient decision support to physicians and the families of children by developing AI for the early detection of *Cerebral Palsy* (CP). To achieve this goal, the DeepInMotion project utilize a large global database of recordings of newborns at a high risk of developing CP, enabling the creation of the next generation of intelligent systems.

4.2.1 Overview of the technology

The XAI system they are working on focuses on prediction accuracy compared to human experts. However, the accuracy of the explanation itself is still being researched and quantified. The direction they are likely to take is providing procedures to test the correctness of the explanation and establish trust in it, rather than aiming for a definitive percentage of correctness.

Regarding the explanation process, they are employing a post-hoc explanation approach. This means that the explanation is obtained after the prediction is made. The developer in the team mentioned the possibility of using classical machine learning methods instead of deep learning, to achieve explainability. The original architecture of the XAI system was developed by a previous PhD student and the architecture primarily utilizes deep learning, which is known for its lack of explainability. However, that model demonstrated similar accuracy to that of human experts, and the model incorporates a larger amount of data compared to the researchers ability to incorporate large amount of data, which enhance confidence in the systems performance.

Overall, the XAI system aims to achieve high prediction accuracy and is exploring ways to achieve both transparency and explainability. The approach involves post-hoc explanation and utilizes deep learning techniques with a focus on incorporating substantial amounts of data for improved reliability.

4.3 Automated assessment of left ventricular function by advanced ultrasound - Case C

Automated assessment of left ventricular function by advanced ultrasound aims to estimate regional function in the left ventricle. This is important because global function, which is described by the overall function of the ventricle, is a significant predictor of various pathologies in cardiac surgery. However, the project aims to address the limitation of global function assessment by focusing on detecting infarcts in specific segments and estimating deformation in the heart wall. At present, the development team is actively working on the implementation of two-dimensional (2D) and three-dimensional (3D) techniques within the XAI system.

Case C has not applied for approval for their XAI system to be recognized as a medical device. They explain that it was easier to operate with patients without seeking such approval. However, they question the legality and how long they can continue in this manner. They clarify that currently, the system is not considered a medical device as it does not directly impact patient care. However, if they intend to make it a medical device, they would need to ensure that it goes through the appropriate approval process.

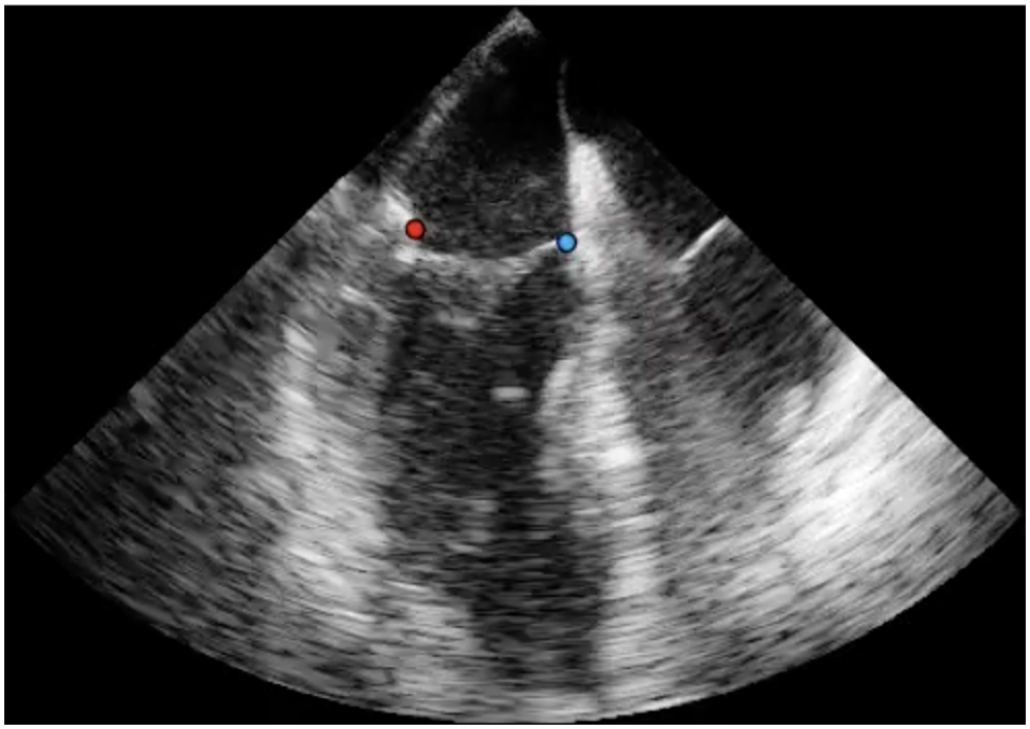


Figure 4: The left ventricle with the two marking points on each side to measure the heartbeat.

4.3.1 Overview of the technology

The XAI system is tasked with assessing the quality of results and filtering out poor outcomes, presenting them to clinicians without delving into specific error explanations. The primary focus is on discarding errors rather than investigating the underlying reasons, as clinicians can identify them based on their knowledge.

Currently, they are developing methods to estimate both 2D and 3D mapsets in the XAI system. In the 2D approach, the system identifies two key points to provide an interpretable explanation to cardiologists. By visualizing these points, clinicians can evaluate anomalies or errors and potentially explain the system's predictions. However, the 3D mapset visualization method is more subjective, relying on individual clinicians to assess the correct placement of points. This process may be challenging and less accurate in a clinical setting, allowing for diverse interpretations. Validating and verifying the system's outputs is more difficult in the 3D approach due to less intuitive visualization and interpretation.

5 Result / Findings

In this section, I will present my findings from the research questionnaire in a structured manner. The first subsection (5.1.1) will address the challenges clinicians face when utilizing AI for medical diagnosis. The second subsection (5.1.2) will present the challenges developers encounter when developing AI systems for medical diagnosis. Finally, the third subsection (5.1.3) will focus on how multidisciplinary cooperation can help mitigate these challenges.

To gather the necessary information for this study, I conducted interviews with clinicians involved in three case studies. This forms a significant portion of the empirical basis for my findings. The rest of the study's empirical basis comprises the interview information collected from system developers.

5.1 Challenges faced by clinicians

5.1.1 Explainability and Transparency

In case A, a sequence of images with a colored image of the prostate is delivered to the radiologists as the output. Then, depending on what color the area is, the color should indicate whether the risk is high or low for prostate cancer. The system should be a tool to aid physicians in diagnosing the patient so that they can utilize it in addition to the traditional methods of patient interpretation. The problem with this system is that it only visualizes the colored areas and the probability of cancer in the colored area. The clinicians still need to know which variables the machine considered while creating the image sequence.

> *" Det er jo noe som gjør det vanskelig å diagnostiser prostata, at det er så stor variasjon mellom folk. Men her er det sånn at, utifra de MR bildene vi allerede tar så skal det verktøy samle all den informasjonen som ligger i de bildene, og ting som ikke er direkte synlige for radiologen visuelt. " - Case A*

In case B, a video is displayed together with points that are visually positioned on top of the child's body in the video. These points will help draw attention to the child's motions, making it more straightforward for the physician to determine whether the points and movements interact and whether the patient or child is in jeopardy of receiving a CP diagnosis. Additionally, the system outputs a number ranging from 0 to 1. The approach evaluates the child as having a high risk of CP if the number is more significant than 0.35. According to clinicians, the points that highlight the body movements give a certain amount of trust in the system. On the other hand, if the clinician disagrees with the system's evaluation or is doubtful, the clinician is unaware of what the system has highlighted in the body motions. Similar to Project A, this will increase mistrust in the clinician.

"Vanskelig å stole på et system når man ikke vet hvilke data som modellen er trent på."
- Case B

Case C's technology will examine the patient's heart by inserting a tube through the patient's mouth. On a computer next to the clinician, the clinician will view an ultra live video in which the heart is visibly beating. The clinician can see two marking points in the video of the heart, which is observed beat by beat. The movements on each stroke are then measured at these two markings, resulting in a millimeter-scale number. A challenge occurs when there is no relationship between millimeters and the heart's health. The illness prevents the heart from beating in unison, but the ultrasound video will not show this. If the millimeter output is a number that reflects a healthy heart, then the millimeter output will be taken to mean that the heart is healthy. Therefore, the system is unaware of this medical condition and cannot depict it in the video or the graph shown in parallel. Serious repercussions and incorrect diagnoses may result from this. It is, therefore, essential that clinicians know such simplifications in advance and which factors are emphasized.

"Ja, det kan være feil likevel. Ja, uten at man legger merke til det. Jo, ja det kan det."
- Case C

These results underline how crucial it is to comprehend how the system generates its output. The clinicians' ability to trust the system's recommendations may be limited by its lack of explicability. Clinicians may need help understanding and trusting the system's output if they are unaware of the underlying information and logic underlying the value. In the diagnostic procedure, where the clinician's experience and judgment are still vital, the system is consequently seen as a supporting tool. Another topic of uncertainty among clinicians is the notion of explainability. This shows that there is a need for better communication and comprehension regarding the interpretability of the system and the variables affecting its output. Concerns regarding using AI systems for medical diagnosis could be reduced, and trust could be increased by improving explainability and providing transparent decision-making procedures.

5.1.2 Bias

In the context of visual representation, it can be problematic if the models are not trained on a diverse range of skin colors and tones. Specific standards must be followed to ensure the system provides accurate classification. For example, in case B, when capturing images, it is essential that a child is not crying and is comfortable. The clinician is also unaware of how the system will behave depending on the type of skin color.

"Det er jo for eksempel det med spedbarn, at barnet skal være våkent, og bare ha det komfortabelt, og ha det ok. Det betyr at hvis for eksempel barnet gråter, eller nesten

halvsover, eller bruker en smukk, så vet jeg at det virker inn på hvordan barnet beveger seg. Det endrer jo de dataene som puttes inn i systemet, og vil jo kunne være en feilkilde med ved klassifisering ut fra AI-algoritmen." - Case B

In case A, the clinician believes that the prostate is quite similar around the world, based on MRI. However, they are still determining if this knowledge is directly transferable. In addition, even though the clinician believes that they prostate is similar worldwide, the prostate itself is quite complex, and even more complex to diagnose with prostate cancer.

"vi tror at prostata er ganske lik verden over. Bedømt bare på MR så er det likt." - Case A

In case C the clinician has noticed that the system tends to favor images with the best quality. It is not always obvious who has good or poor quality images; if the clinician is unaware of this, it can be very harmful.

"Altså, jeg har hørt om det der problemet med intelligens og modeller og sånt, men jeg har ikke helt skjønt hvordan det skal være." - Case C

"Altså, den favoriserer jo de med best bildekvalitet. Ja. Og det er ikke alltid selvfølgelig hvem som har bra bilder og hvem som har dårlige bilder." - Case C

When it comes to trusting the data set on which the AI system is trained, the clinician from Case B raises concerns. From talking with the developer, the clinician knows that the dataset includes of videos of adults executing an action such as drinking water. This breeds skepticism in the system since the clinician I interviewed cannot comprehend how such a data set may reflect children's movements and hence contribute to the diagnosis of children with CP.

"Da blir det jo fort spørsmål om om AI-algoritmer har vært trent nok på alle de eventualiteter av data som er pasientene der ute. Da kjenner jeg at jeg blir mye mer skeptisk og usikker." - Case B

5.1.3 Trust

All the systems aim to contribute to quality improvement in medical diagnosis. However, one issue mentioned by all clinicians is that, ultimately, the clinician must make the decision and diagnose the patient. In this process, the clinician's training and expertise are essential. The clinician from case A says that unthinkingly depending on the system without the proper training or knowledge to evaluate the outcomes can have disastrous results and result in misdiagnosis. The contrast

between radiologists who examine two patients daily in a large hospital with a high patient volume and radiologists who only examine two patients per month in a hospital with fewer patients is brought up by another clinician. The latter group does not regularly interpret the outputs of AI systems.

> *"Det er også et problem, at det er veldig opp til tolkningen til hver radiolog. Og hvor god utdanningen har jeg når jeg sitter å tolker det, og hvor ofte gjør jeg det." - Case A*

The clinician from case C expresses that confusion can arise when they disagree with the system's results. They emphasize that a user's dependency on the system is heavily influenced by their knowledge and understanding of their field. This means that the explanations provided by the system are subject to individual interpretations based on the user's existing knowledge. Consequently, different interpretations of the system's output can arise.

> *" Så det er også mange årsaker til det som ikke bedre nødvendigvis har med bra hjertepumpe. Hvor mye hjertet beveger seg, det måler millimeteret hjertet bevege seg" - Case C*

The clinician's observations in case B show difficulty building confidence in the created system, a critical issue covered in this master's thesis.In terms of the applicability of their AI system to broader contexts, the clinician raises a critical query, asking whether it was trained on movement data from infants with observed risk factors or on completely healthy kids. This mismatch raises concerns about the system's accuracy in determining typical children. The physician highlights several crucial elements that help foster confidence in the success of the system in order to solve this challenge. He underlines the importance of patient selection to ensure the correct population gets medical attention using the system. The clinician also emphasizes the importance of accurate movement video recording and adherence to predetermined standards, which serve as the basis of the diagnostic procedure. By following these guidelines, the system may gain confidence from the clinicians. Therefore, the clinician understands that building confidence depends on elements like sensible patient selection, specific data collecting, and adherence to predetermined processes, all of which contribute to the system's overall trust.

> *"Hvis jeg skal ha tillit til det systemet jeg selv har vært med på å utvikle. Så tror jeg faktisk at det handler veldig mye om hvem som er pasienten du bruker det på. At det er riktig gruppe, at videopptak av bevegelsene er gjort på riktig måte, at det er en del standarder som følges, som er grunnlaget for selve undersøkelsen." - Case B*

> *"Det er at vi har trent algoritmen vår på bevegelsesdata for spedbarn, som da har risikofaktorer. Og et stort spørsmål vi sitter på nå, det er, kan vi bruke den AI-algoritmen til å undersøke helt friske vanlige barn? For det har jo ikke inngått i vårt*

treningsgrunnlag av algoritmer. Så da er spørsmålet, kan vi ha tillit til at systemet virker på helt vanlige barn?" - Project B

In case A, clinicians highlighted the limited availability of resources for understanding the output of the AI system. In such cases, their only recourse is to seek assistance from colleagues or rely on their own expertise to diagnose using conventional methods. The confirmation of this issue in the other two projects further emphasizes its significance.

"I realiteten hvis jeg er i tvil i verktøyet så har jeg ingen å spørre." - Case A

5.2 Challenges Faced by Developers

5.2.1 Bias

The data collected when conducting interviews with developers involved in the three projects, included important findings regarding challenges and considerations in the context of data sources, standards, and potential model biases.

The developers involved in case A emphasized the importance of considering standardization and data diversity when applying AI models. They highlighted that the available data primarily consisted of data on a white population. However, they acknowledged that prostate anatomy and characteristics can vary based on ethnic backgrounds. As a result, there were concerns about how the model would perform in different regions, such as Ghana or South Africa or predominantly black communities in the United States. This raised questions about the generalizability and effectiveness of the AI model across diverse populations.

"This data is largely I would say based on, white population data sets. (...) If you take this model to Ghana South Africa or to black dominant residence in the US, how will it perform, it is a question that we are here to see." - Case A

Furthermore, the developers noted that the international standardization of scanning prostate cancer with MRI primarily occurs in Europe, with limited implementation in the United States. The accessibility and affordability of MRI scans were identified as crucial factors influencing their prevalence. While MRI scans are widely conducted in Europe due to public healthcare systems, their utilization is more limited in other countries. This discrepancy in availability and image quality posed challenges in creating a representative and diverse data set. The data set used in Project A mainly relied on the 3 Tesla data set, which may not fully capture the variations and characteristics of prostate cancer patients. Due to their greater picture quality, 3 Tesla MRI scans are widely used in Europe, which has improved diagnostic capabilities. However, the prohibitive costs of 3 Tesla scanners prevent their widespread use and accessibility, which reduces the Project

A data set's diversity. Instead, the prevalence of 1.5 Tesla scanners, which is primarly due to their price, introduces a tradeoff with image quality and may have an adverse effect on the precision of the medical diagnosis system. Additionally, we saw differences in the use of MRI scans for prostate patients, particularly in nations without universal healthcare because of the high prices that prevent widespread use. These findings highlight the importance of taking into account both technological and economic issues when creating data sets for MRI scan-based medical diagnosis.

> *"International standardisation of scanning prostate cancer with MRI is actually applied mainly in Europe, even in the US it is applied, but on a very small scale, because in Europe the healthcare systems are public, which make it free or in a way, cheap to do the scanning. (...) MRI scan is super expensive. So, I would say in the greatest majorities of the countries, prostate cancer patients will not go for an MRI scan." - Case A*

The current developer from Project B recognizes bias in their data set because they are currently using the a data set that concentrates on adults performing activities like sipping water or waving their hands, to test their XAI techniques. The developer however acknowledge that using the same techniques on babies who behave differently serves a different goal and can have different applicability. Identification of adult behavior is not the current goal of their research, it is identification of an infant's behaviour. That is why the developer soon will broaden the application of their techniques to determine whether or not infants have CP. The developer underlines that the shift from examining adult activities to evaluating baby motions can be a significant transition in their research and may cause some unexpected challenges.

> *"I'm using this data set now to test my XAI methods. But it is that the subjects are adults performing actions. And then I want to apply the same XAI methods to babies. Performing a different action. And it has a different application. So in my research now, it's just to identify what a person is doing. Like drinking water or hand-waving. But later on I will apply it for babies and see if they have CP or not. So Lars, for him, it seems like it's a big difference. " - Case B*

In the context of case C, the developer claims that the data set they are working with has little diversity. This can be due to a number of reasons. First of all, the data set is derived from a focused research project in which interventions are restricted to a single cohort under evaluation. Second, the data set only includes patients who were subjected to a single type of ultrasound system, eliminating individuals who were treated with other ultrasound systems. As a result, the data set has an inherent bias, which the developer is aware of, but contradictory to their original goal. This describes the present state of affairs with relation to the data set.

> *"Dataen er pasientdata fra St.Olavs, så det er jo en spesifikk forskningsgruppe. Det er også en spesifikk ultralyd-leverandør. Vi har jo ikke lyst til å ha nettverk som skiller på*

minoriteter. Men det systemet er det som er mest utsatt for det, for det er ikke store variasjoner i den type data i vårt tilfelle." - Case C

5.2.2 Explainability and Transparency

During the interviews conducted with developers from three distinct projects, referred to as case A, case B, and case C, valuable insights were gathered regarding the importance of visualizing AI decisions, the challenge of validating AI models, and the distinction between explainability and transparency.

In case C, developers acknowledged that the current approach relies heavily on individual clinicians to interpret the AI-generated results and understand the underlying connections. However, they recognized that it might be ambitious to expect clinicians to grasp and visualize the relationships easily. The visual representation of AI decisions could be more straightforward because now it is making it challenging to validate whether the interpretations made by clinicians align with the intended understanding of the AI system's output.

"Ja, for akkurat nå blir det jo veldig opp til hver kliniker selvfølgelig, å se sammenhengen. (...) Mulig at det er litt ambisiøst å forvente av en kliniker." - Case C

The ability to observe and comprehend AI judgments at any moment in case A was emphasized by developers. The model was created with transparency in mind, allowing developers and clinicians to understand why certain decisions were made. They aimed to connect the model's conclusions, how the model weighted features in the algorithm, and clinical features that radiologists were familiar with. By including 100 or more mathematical elements with transparent operations, the developers aimed to provide clinicians with a clear understanding of the model's decision-making process. They discovered that rather than relying entirely on the outcome, clinicians frequently try to understand the explanation behind the results. In practice, "explainability" and "transparency" were used interchangeably; however, the developers distinguished transparency as the knowledge of how the model arrived at its results.

"So I think the word transparency and explainability are kind of the same." - Case A

In case C, challenges were identified in visualizing the relationships and validating AI decisions due to the complexity of the visualization process. In contrast, case A emphasized the importance of visualizing decisions and providing transparency in the model's decision-making process. Developers aimed to establish a clear understanding of the relationship between AI decisions and clinical features, enabling clinicians to comprehend the reasoning behind the results.

"Det er ikke like lett å visualisere for klinikere, så det er ikke like lett heller å validere om man etterhvert gjort det riktig eller ikke." - Case C

"So at any point in time, we can visualize or we can go into the model to see why decisions were made. And we can link them to some kind of clinical features in a way to put it that the radiologists can also relate to." - Case A

In case C, the developers acknowledged that there are various simplifications and situational factors that can lead to estimation errors. However, they emphasized the necessity of involving a clinician in the assessment process regardless of these challenges. The developers stressed that a clinician's expertise is crucial to evaluating such situations, underscoring the importance of human judgment and decision-making even in the presence of AI-based estimations.

"Endeproduktet ønsker å gi en heads up til anestesiologer og kardeloger om at det er redusert hjertepunksjon, i håp om å kunne interagere tidlig hvis det er noe patologi. Så kan man jo potensielt gi smittesummefeil som fører til at man enten overser svikt i hjertepunksjonen der det er svikt, eller motsatt. Det kan jo være alvorlig det." - Case C

In case A, explainability is achieved using a mathematical approach with radiomoic features. These features comprise approximately 100 or more handcrafted mathematical features. One of the developers in case A is actively working on simplifying the explanation process further. The goal is to provide doctors with easily understandable terms related to the features used in the model. If the AI model suggests a high likelihood of cancer in a specific area and the clinician disagrees or desires further clarification, the doctor can explore the decision-making process. By examining the features and their relevance, the clinician can gain insights into how the decision was reached. The AI system functions as an assistant tool, providing valuable information, but the final clinical decision regarding the presence of cancer rests with the radiologist or medical doctor.

"Explainability is like in our case, we use something called radimoic features, which is basically a mathematical-based handcrafted features. It's kind of 100 or something, 107 features." - Case A

"If the model proposes this area to be high cancer likelihood and for some reason you don't agree or you want to know for that, then that's where you go for that to see how the decision came about, what features you're looking at, and then you go in for that." - Case A

5.3 Mitigating the current challenges

5.3.1 Team Collaboration

Various thoughts regarding transparency, explainability, and collaboration emerged throughout the interviews with developers from case A, case B, and case C. The developers discussed their experiences, strategies, and methodologies when incorporating these ideas into their products.

In case A, the developers emphasized the involvement of two radiologists who played a significant role in the project. The radiologists provide valuable input on transparency by assisting in fine-tuning the system and offering feedback on the obtained results. This collaboration extends beyond email communication, as the developers regularly consult with the radiologists in person. Weekly meetings are held, and one or more radiologists participate whenever possible to facilitate effective communication. Additionally, the developers have planned user guidelines, which will be shared with the radiologists through workshops and one-on-one training sessions. This comprehensive approach aims to ensure transparency and align the AI system with the radiologists' expertise and requirements.

> *"So whatever results we get, we update them, and they give us feedback. " - Case A*

Case B's developers recognized the importance of communicating the concept of XAI to clinicians involved in the project. They go into extensive detail about the AI system's strengths and weaknesses. When it comes to the usage of AI in their field of practice, clinicians usually have distinct visualization preferences and expectations. The developers underline the need of discussing these design decisions with the the clinicians in order to foster mutual understanding. To facilitate communication and team updates, both technical and non-technical employees attend weekly meetings. These sessions allow for the presentation of relevant research articles as well as the resolution of team questions and issues. Additionally, the developers utilize a Teams group as a platform for quick and convenient communication, enabling anyone to seek assistance or raise inquiries.

> *"We have to explain to him like what is explainable AI? What is it capable of? And what are its limitations. (...). For example: because clinicians, they have some ideal way to visualise AI and how they can use it for their work. But it's not necessarily possible or applicable. So we explain that to the clinician. " - Case B*

The developers of case C also emphasized the significance of ongoing communication and informal debates about transparency and XAI. While these subjects may not be specifically addressed in formal agendas, they are often addressed and discussed in a general sense during project-related meetings. The developers noted that the technical nature of XAI frequently required the involvement of supervisors or other experts in the field to provide more thorough insights. The concept

of explainability is developed as the project progresses, making it challenging to adhere to clear guidelines before they become apparent.

> *"Ikke spesifikt. Det har kanskje ikke vært punkter på agendaen, men punkter man har snakket rundt." - Case C*

> *"Vi snakket litt om det, men ikke i dybden, for teamet vårt består stort sett av klinikere. Og mye av samtalene om explainable AI er kanskje mer tekniske, og det er mer naturlig å ta med veileder." - Case C*

5.3.2 Relation to EU MDR

In case C, the developers acknowledged that their team mainly consists of clinicians, This has influenced in how much detail they discuss and understand the concept of XAI. While they touched upon the topic, it was not extensively explored. The developer said that conversations regarding XAI often tend to be more technical, making it more natural to involve his supervisor or expert in the field to provide further insights. The developer also said that their current method of AI development has significant black box characteristics, indicating limited transparency in the system's inner workings. Although the developer acknowledged the importance of explainability and transparency, the focus thus far has been more on the technical aspects of their methodology. The developers mentioned that while specific agenda points on explainability may have yet to be specifically included in their discussions, the topic has been informally addressed and discussed more broadly. These conversations shaped their understanding and approach to the concept of XAI, even if it was not a formalized aspect of their project. Furthermore, the developer noted that the teams approach to explainability and transparency in their project was vague and evolved during the development process of the system. The developer noted the challenge of following to the EU MDR standards when the terms transparency and explainability do not yet have established definitions.

> *"Det er vagt, og så har det blitt formulert mens vi har drevet med utvikling. Så det er vanskelig å følge klare retningslinjer før det eksisterer. " - Case C*

Transparency has surfaced as a critical obstacle in case A. The lack of a clear description of the level of transparency to be applied has complicated the project. This vagueness raises concerns about potential differences in transparency levels between initiatives. As a result, the project team has prioritized the deployment of transparency and explainability to guarantee that consumers have a clear grasp of the technology. Explainability and transparency are relatively interchangeable notions in the developer's mind. Their major goal is to let users to watch and comprehend the behavior of technology, regardless of whether it is referred to as explainability or transparency.

"they say transparent, but there's no clear. It's not clear at all. You've got different people claim different levels of transparency as transparent, especially when you're talking about deep learning vs classical machine learning." - Case A

The developer in case B similarly did not thoroughly examine the EU MDR and hence did not base the built system on the specified standards and guidelines that are mentioned in the regulation. Despite this, the significance of regulation is underlined, as is the importance of having explainability and openness in the project's structure.

"I don't think we have really checked the MDR and design our project based on MDR. But I think we kind of understand like, OK, we need some sort of transparency, explainability in our model." - Case B

6 Discussion

The discussion section will provide an extensive review of the findings and be structured according to the research questions. The first subsection (6.1) will discuss and interpret the challenges developers and clinicians face in the context of AI systems, with a particular focus on bias and the need for explainability and transparency. The second subsection (6.2) will discuss how the challenges can be mitigated.

6.1 Challenges developers and clinicians face utilizing AI systems

6.1.1 Explainability and transparency

The challenges clinicians face when using AI systems for medical diagnosis are many. A significant concern is that the system lacks explainability and transparency when utilizing the system. In Project A, radiologists are presented with images containing a colored representation of the prostate, indicating the probability of prostate cancer in different areas. However, the system does not provide information about the variables considered when generating the image sequence. Clinicians know that some variables are considered and that, to some extent, there are different variables from patient to patient because the prostate differs from man to man. This, therefore, leads to clinicians needing to be informed of which functions or patterns the machine has considered to reach its conclusions. This lack of transparency in the system's decision-making process hinders clinicians' ability to fully understand and trust the system's recommendations. Therefore, the clinician must use their prior knowledge to interpret the images, although the system was initially supposed to lead to quality improvement and higher accuracy when making a diagnosis.

Similarly, in Project B, clinicians are presented with a video of a child's body with points placed on specific parts. These points highlight the child's movements and should help diagnose CP precisely because the children's movements should be highlighted for the clinician. However, suppose the clinician disagrees with the system's evaluation or is unsure why a body part and movement are highlighted. In that case, the clinician cannot backtrack the specific features that influenced the system's outcome. The transparency of the system's decision-making process limits the clinician's ability to trust and rely on the recommendations.

Project C examines a patient's heart using an ultrasound video. Clinicians observe two marking points in the video to measure the heart's movements. However, the system's output in millimeters fails to capture some critical conditions that affect heart health, and therefore they are not visually evident in the ultrasound video. This is because the system has made some simplifications and has not included all conditions yet. The system's inability to represent these conditions gives cause for concern about incorrect diagnoses and potentially severe consequences for the patient because it will then be the clinician's knowledge that will impact the diagnosis, which can vary widely

from clinician to clinician. In addition, if the system's strengths and weaknesses have not been communicated well enough between the developer and clinician, the clinician may believe that all conditions have been included in the system and thus have complete trust, which can have fatal consequences. Once again, the lack of transparency regarding the system's limitations and the variables assessed to generate output minimizes clinician confidence and limits understanding.

The clinicians' need for explanation and transparency is rooted in the importance of their expertise and judgment in the diagnostic process. While AI systems can improve medical diagnosis, clinicians play a critical role in interpreting and applying the system's results to a patient's individualized conditions. With a clear understanding of the system's decision-making process, clinicians can trust the recommendations fully and rely more on their clinical judgment. Improving explainability and transparent decision-making procedures can help address concerns, increase trust, and promote effective collaboration between clinicians and AI systems in medical diagnosis.

Developers are also tasked with creating AI models clinicians can interpret and explain. Despite the solid predictive capabilities of deep learning models, they are often considered black boxes, making it challenging for clinicians to understand and trust the results of these models. Transparent models are essential for the clinic to understand the factors influencing the system's predictions and effectively validate the recommendations. Finding a balance between model complexity and transparency is a crucial challenge for developers. They need to explore techniques and methods that improve the transparency of AI models so that clinicians can understand the system's decision-making process.

6.1.2 Bias

Bias poses another challenge when using AI systems for medical diagnosis. In the framework of visual representation, such as in Project B, training the system on various skin colors and tones is essential to ensure accurate classification from the system. Lack of diversity can lead to incorrect classification and inaccuracies in diagnoses. In addition, factors such as the child's comfort and behavior during image recording can affect the system's output. In Project B, clinicians note that if a child cries, is half-asleep, or uses a pacifier during imaging, it can affect the child's movements and, consequently, the data fed into the system. This introduces potential errors and bias in the AI algorithm's classification.

Furthermore, in projects A and C, the clinicians recognize the importance of considering diversity in the data sets when using AI models. Like heart health, Prostate anatomy and characteristics may vary across ethnic backgrounds and populations. However, the data sets used to train these models consist mainly of data from white populations. This raises concerns about these models' broader relevance and efficacy across different demographic populations. Clinicians question whether these models will work together.

Access to extensive and varied medical data sets is required to train accurate and reliable AI models. However, developers often need help obtaining such data sets due to limitations concerning the use of different equipment worldwide, privacy concerns, and strict regulations. Limited access to different data can hinder the development of robust AI models that effectively provide good results across different populations and capture rare conditions. In addition, data quality for training reliable models is an issue. Incomplete, incorrect or inconsistent data can introduce biases and affect the performance of the AI systems. Developers must therefore put more effort into obtaining high-quality medical data to ensure its accuracy and integrity.

Developers must also navigate the ethical considerations surrounding AI systems in medical diagnosis. Protection of patient privacy and data security is of importance, especially when it comes to sensitive medical information. Developers must implement robust privacy mechanisms to secure patient data and gurantee compliance with relevant regulations and standards.

Furthermore, addressing biases in AI systems is a critical challenge. Developers must actively identify and mitigate biases concerning the data quality discussed above. They must develop strategies to train models on different data sets that accurately represent the population they aim to serve. By involving clinicians and domain experts in the development process, developers can gain valuable insight into potential biases and collaborate to minimize their impact.

6.2 Mitigating the current challenges

To overcome the challenges mentioned above, collaboration between clinicians and developers is a challenge. This section discusses strategies and the approaches identified in the interviews that aim to mitigate the challenges related to transparency, explainability, and collaboration in developing AI systems for medical diagnosis. The findings provide insight into the developers' actions in the three projects to meet these challenges.

6.2.1 Team Collaboration

The developers in Project A emphasized how important it was to collaborate with radiologists throughout the development process. With the involvement of radiologists, the developers can fine-tune the AI system and get feedback on the results. This collaboration extends beyond e-mail communication, regular personal consultations, and weekly meetings. By actively engaging with radiologists, the aim should be to ensure transparency and adapt the AI system to the expertise and requirements of the radiologists. The planned user guidelines that the developers have gained knowledge of through the EU MDR, which will be shared through workshops and one-to-one training sessions, further contribute to transparent communication and collaboration.

In Project B, the developers recognized the need to educate clinicians about explainable AI. They

provided detailed explanations about the technical capabilities and limitations of the AI system, thereby promoting a shared understanding between the technical team and the clinicians. Weekly meetings were organized to adapt smooth communication and updates among team members, and a Teams group was established as a platform for practical communication and assistance.

Project C emphasized ongoing communication and informal discussions related to transparency and explainable AI. Although these topics were not explicitly included in formal agendas, they were often discussed during project-related conversations. Developers acknowledged that the technical nature of explainable AI often requires the involvement of supervisors or external experts to provide extensive insight. As the project progressed, the concept of explainability was refined, which gave rise to challenges in following the required guidelines that still needed to be established and sufficiently prepared and explained in the EU MDR document.

These findings show different approaches to collaboration to meet the challenges of transparency and explainability. Project A emphasizes the involvement of domain experts (radiologists) throughout the development process, while Project B emphasizes educating clinicians about the AI system. Project C uses continuous communication and informal discussions to refine the concept of explainability. These strategies facilitate the effective implementation of transparency and explainability, thus trying to meet the needs and expectations of various stakeholders involved in the projects.

6.2.2 Relation to EU MDR

The developers of Project C recognized the importance of focusing on technological concerns rather than a thorough investigation of legislation such as the EU's Medical Device Regulation (MDR). Even if specific standards were difficult to discern, they acknowledged the need for transparency and explainability in their model development. They discussed the challenge of adhering to criteria that have yet to be fully defined, so their view of explainability evolved during development.

Project A found a challenge in defining transparency based on what they read and heard from people who perceived varying amounts of transparency. Although they had yet to thoroughly investigate legislation and the EU MDR guidelines, the developers concluded that transparency and explainability are inextricably linked in the technical aspect of their project because a transparent system is likewise explainable.

Project B did not mention the EU MDR directly but acknowledged the importance of transparency and explanation in constructing its approach.

These findings indicate that developers have various degrees of understanding and devotion to transparency and explanatory ideas and concern for existing and proposed legal frameworks, such as the EU MDR. While some projects have concentrated on technical concerns and informal talks about these topics, others have focused on the perceived need for transparency and regard it as

similar to explainability. The findings highlight the value of further investigation, clarification, and consistent interpretation of transparency and explanation in creating XAI systems. They also emphasize the importance of developers staying educated and adhering to current legislation and guidelines, such as the EU MDR.

7 Conclusion

The discussion section focuses on the challenges developers and clinicians face in connection to the use of AI systems for medical diagnosis. The challenges discussed is the need for clarity and openness, the presence of biases, data availability and quality, model interpretability, ethical considerations, and limitations inherent in studies.

Clinicians need help with using AI systems due to the lack of explainability and transparency in decision-making. The opacity of the system's algorithms and the absence of information about the variables being assessed hinder clinicians' trust and understanding. In addition, biases in AI systems can lead to misclassifications and inaccuracies, especially when the under-representation of different populations is considered. Developers face data availability and quality challenges, model interpretability, and ethical considerations. Limited access to different datasets and high-quality data hinders the development of robust AI models. Interpretable models that can be easily understood by clinicians, who are the reason, pose a challenge to the complexity of deep learning models. Ethical considerations, such as privacy and bias, must also be addressed during development. To mitigate these challenges is a collaboration between clinicians and developers. Strategies such as teamwork, educating clinicians about AI systems, and compliance with regulations such as the EU MDR have been identified as some practical approaches. These strategies promote transparency, explainability, and understanding between stakeholders involved in developing and implementing AI systems for medical diagnosis.

The implications of the findings highlight the importance of addressing the challenges discussed to improve the effectiveness and reliability of AI systems for medical diagnosis. Improving transparency, explainability, and collaboration between developers and clinicians can improve outcomes and promote confidence in AI recommendations. Moreover, addressing biases and ensuring the representativeness of different populations in training data sets are crucial steps in achieving fair AI systems. Overall, this study contributes to answering my research question:

Why or how do clinicians face challenges in the utilization of AI systems for medical diagnosis, and how will multidisciplinary cooperation contribute to mitigating these challenges?

In summary, this study enhances our knowledge of the challenges involved in developing and utilizing AI systems for medical diagnosis and suggests potential approaches to address them. By addressing these challenges, the healthcare industry can make crucial improvements in utilizing AI technology.

7.1 Limitations and Further Research

This study has certain limitations. It is essential to acknowledge that time restrictions prevented the development of a fully functional prototype of the XAI systems for clinical testing and application in the three cases. As a result, rather of depending on actual results and reactions gained from fully constructed prototypes, this master thesis primarily focus on analyzing clinical expectations and predicted reactions based on their understanding of the system. This limitation highlights the emphasis on clinician viewpoints and insights regarding the system's potential benefits and limitations, rather than giving actual evidence based on real-world prototype use.

The interviews were conducted with few participants, representing three projects. Expanding the scope of research by including a wider range of experiences and perspectives is crucial to ensure that the findings presented in this master's thesis are not limited. Additionally, it is important to note that all the projects investigated in this study are from Norway. This has implications for the data sets, including the representation of individuals and the equipment used, such as using the same type of MRI machine, for example. Future research should integrate a larger sample size and explore multiple perspectives to gain a more comprehensive understanding of the challenges developers and clinicians face in the context of AI systems for medical diagnosis.

Further research can also focus on involving clinicians in the design process of the system. This would include an process where the developers and clinicians collaborate closely, allowing clinicians to provide feedback on the system, iteratively while the system is developed. Such an approach can contribute to increasing the value of a XAI system in the healthcare sector if clinicians can fully harness the strengths offered by the system. Achieving a positive user experience is crucial in fulfilling this goal.

Bibliography

Alex, Z. (Aug. 2022). *Developing Trust in Black Box AI: Explainability and Beyond — Wilson Center*. URL: https://www.wilsoncenter.org/blog-post/developing-trust-black-box-ai-explainability-and-beyond.

Amann, Julia et al. (2020). 'Explainability for artificial intelligence in healthcare: a multidisciplinary perspective'. In: *BMC Medical Informatics and Decision Making* 121, pp. 1–9.

Arrieta, A. B. et al. (2020). 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI'. In: *http://www.elsevier.com/locate/inffus* 34.

Beckers, R., Z. Kwade and F. Zanca c (2021). 'The EU medical device regulation: Implications for artificial intelligence-based medical device software in medical physics'. In: *Physica Medica* 82, pp. 1–8.

Beil, JMichael et al. (2019). 'Ethical considerations about artificial intelligence for prognostication in intensive care'. In: *Intensive Care Medicine Experimental* 7, pp. 1–13.

Bharati, Subrato and M. Rubaiyat Hossain Mondal (2022). 'A review on Explainable Artificial Intelligence for Healthcare: Why, How, and When?' In: *IEEE Transactions* 31, pp. 538–584.

Bianchini, Elisabetta and Christopher Clemens Mayer (2022). 'Medical Device Regulation: Should We Care About It?' In: *Artery Research* 28, pp. 55–60.

Borys, Katarzyna et al. (2023). 'Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches'. In: *European Journal of Radiology* 162.

Campos Aranovich, atiana de and Rita Matulionyte (2022). 'Ensuring AI explainability in healthcare: problems and possible policy solutions'. In: *https://doi.org/10.1080/13600834.2022.2146395*.

Chu, Lan (n.d.). *Model Explainability - SHAP vs. LIME vs. Permutation Feature Importance — by Lan Chu — Towards AI*. URL: https://pub.towardsai.net/model-explainability-shap-vs-lime-vs-permutation-feature-importance-98484efba066.

Defensiv medisin og hvordan det påvirker helsekostnader", url = "https://no.approby.com/defensiv-medisin/a meeting (n.d.).

EU (n.d.[a]). *About the revision*. URL: https://health.ec.europa.eu/medical-devices-sector/new-regulations_en.

— (n.d.[b]). *Sectorial challenges*. URL: https://health.ec.europa.eu/medical-devices-sector/overview_en.

Grote, Thomas and Philipp Berens (2020). 'On the ethics of algorithmic decision-making in healthcare'. In: *Journal of Medical Ethics* 46, pp. 205–211.

He, Jianxing et al. (2019). 'The practical implementation of artificial intelligence technologies in medicine'. In: *Nature Medicine* 25, pp. 30–36.

Huusko, Juhamatti, Ulla Mari Kinnunen and Kaija Saranto (Dec. 2023). 'Medical device regulation (MDR) in health technology enterprises - perspectives of managers and regulatory professionals'. In: *BMC health services research* 23 (1), p. 310. ISSN: 14726963. DOI: 10.1186/S12913-023-

09316-8/TABLES/8. URL: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-023-09316-8%20http://creativecommons.org/publicdomain/zero/1.0/.

Jakhar, D. and I. Kaur (Jan. 2020). 'Artificial intelligence, machine learning and deep learning: definitions and differences'. In: *Clinical and Experimental Dermatology* 45 (1), pp. 131–132. ISSN: 0307-6938. DOI: 10.1111/CED.14029. URL: https://dx.doi.org/10.1111/ced.14029.

Kumar, Vivekanandan S. and David Boulanger (2021). 'Automated Essay Scoring and the Deep Learning Black Box: How Are Rubric Scores Determined?' In: *International Journal of Artificial Intelligence in Education* 31, pp. 538–584.

Kumar, Yogesh et al. (July 2023). 'Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda'. In: *Journal of Ambient Intelligence and Humanized Computing* 14 (7), p. 8459. ISSN: 18685145. DOI: 10.1007/S12652-021-03612-Z. URL: /pmc/articles/PMC8754556/%20/pmc/articles/PMC8754556/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8754556/.

Martinho, Andreia, Maarten Kroesen and Caspar Chorus (2021). 'Ensuring AI explainability in healthcare: problems and possible policy solutions'. In: *Artificial Intelligence in Medicine* 121.

MD, Stefano A. Bini (2018). 'Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?' In: *The Journal of Arthroplasty* 33, pp. 2358–2361.

Minh, Dang et al. (Nov. 2021). 'Explainable artificial intelligence: a comprehensive review'. In: *Artificial Intelligence Review 2021 55:5* 55 (5), pp. 3503–3568. ISSN: 1573-7462. DOI: 10.1007/S10462-021-10088-Y. URL: https://link.springer.com/article/10.1007/s10462-021-10088-y.

Miotto, Riccardo et al. (Nov. 2018). 'Deep learning for healthcare: review, opportunities and challenges'. In: *Briefings in Bioinformatics* 19 (6), pp. 1236–1246. ISSN: 1467-5463. DOI: 10.1093/BIB/BBX044. URL: https://dx.doi.org/10.1093/bib/bbx044.

Miró-Nicolau, M., G. Moyà-Alcover and A. Jaume-i-Capó (2022). 'Evaluating Explainable Artificial Intelligence for X-ray Image Analysis'. In: *Applied Sciences 2022, Vol. 12, Page 4459* 12, p. 4459.

Niemiec, Emilia (n.d.). 'Will the EU Medical Device Regulation help to improve the safety and performance of medical AI devices?' In: *Digital Health* 8 (), pp. 1–8. DOI: 10.1177/20552076221089079. URL: https://us.sagepub.com/en-us/nam/open-access-at-sage.

Oates, Briony J., Marie Griffiths and Rachel Mclean (2022). *Researching information systems and computing*. SAGE Publications, Inc.

Payrovnaziri, Seyedeh Neelufar et al. (July 2020). 'Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review'. In: *Journal of the American Medical Informatics Association* 27 (7), pp. 1173–1185. ISSN: 1527974X. DOI: 10.1093/JAMIA/OCAA053. URL: https://academic.oup.com/jamia/article/27/7/1173/5838471.

Petry, Lucas May et al. (2020). 'Challenges in Vessel Behavior and Anomaly Detection: From Classical Machine Learning to Deep Learning'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*

12109 LNAI, pp. 401–407. ISSN: 16113349. DOI: 10.1007/978-3-030-47358-7_41/COVER. URL: https://link.springer.com/chapter/10.1007/978-3-030-47358-7_41.

Portal, The medical (n.d.). *Medical Ethics - The Four Pillars Explained - The Medic Portal*. URL: https://www.themedicportal.com/application-guide/medical-school-interview/medical-ethics/.

Rasheed, Khansa et al. (2022). 'Explainable, trustworthy, and ethical machine learning for healthcare: A survey'. In: *Computers in Biology and Medicine* 149.

Schwendicke, F., W. Samek and J. Krois (2020). 'Artificial Intelligence in Dentistry: Chances and Challenges'. In: *https://doi.org/10.1177/0022034520915714* 99, pp. 769–774.

Tomaszewski, Michal R. and Robert J. Gillies (Mar. 2021). 'The biological meaning of radiomic features'. In: *Radiology* 298 (3), pp. 505–516. ISSN: 15271315. DOI: 10.1148/RADIOL.2021202553/ASSET/IMAGES/LARGE/RADIOL.2021202553.TBL2.JPEG. URL: https://pubs.rsna.org/doi/10.1148/radiol.2021202553.

Vasiljeva, Ksenija, Bernard H. van Duren and Hemant Pandit (2020). 'Changing Device Regulations in the European Union: Impact on Research, Innovation and Clinical Practice'. In: *Indian Journal of Orthopaedics volume* 54, pp. 123–129.

Weerts, Hilde J. P., Werner van Ipenburg and Mykola Pechenizkiy (2019). 'A Human-Grounded Evaluation of SHAP for Alert Processing'. In.

Wilkinson, Beata and Robert van Boxtel (2020). 'The Medical Device Regulation of the European Union Intensifies Focus on Clinical Benefits of Devices'. In: *Therapeutic Innovation Regulatory Science volume* 54, pp. 613–617.

Xue, Hui et al. (n.d.). 'Clinician's guide to trustworthy and responsible artificial intelligence in cardiovascular imaging'. In: (). URL: https://future-ai.eu/.

Yin, Robert K. (2017). *Case Study Research and Applications: Design and Methods*. SAGE Publications, Inc. ISBN: 9781506336152.

Appendix

A Interviewguide for the clinicians

Step 1: Introduction

- Can you state your name and work title?
- Can you describe what you do in your work?

Step 2: Feedback on system

- Does the system give good enough guidens / information when interpreting the output? (Transferability / Explainability)
- Does the system give the opportunity to see what other recommendations could be given if the inputs were different? (Transferability)
- Where does the system provide explanation? Do you have any examples? (Transparency)
- Do you need to contact other medical professionals or other external information to understand the information in front of you? (Explainability)
  - What other information are you missing?
- Is it clear to you how/why the diagnosis is given? (Explainability)
- Does the system give information regarding any inaccuracy in the result? (Explainability)
- Do you trust that the output/ recommendations of the system is correct? Do you agree with the output? (Transparency)
- Do you understand how to use the system? (Explainability)
- Do the system give possibilities for help, if it is not understandable? (Explainability)
- Do you think that you could've made the same recommendation faster or in the same pace as the system?
- Does the explanation contain personalised arguments for the recommended diagnosis?(Transferability)
- Do you experience that the system favours specific demographics groups? (Bias)
- Do you experience that the system reflect individuals in the society? (Bias)

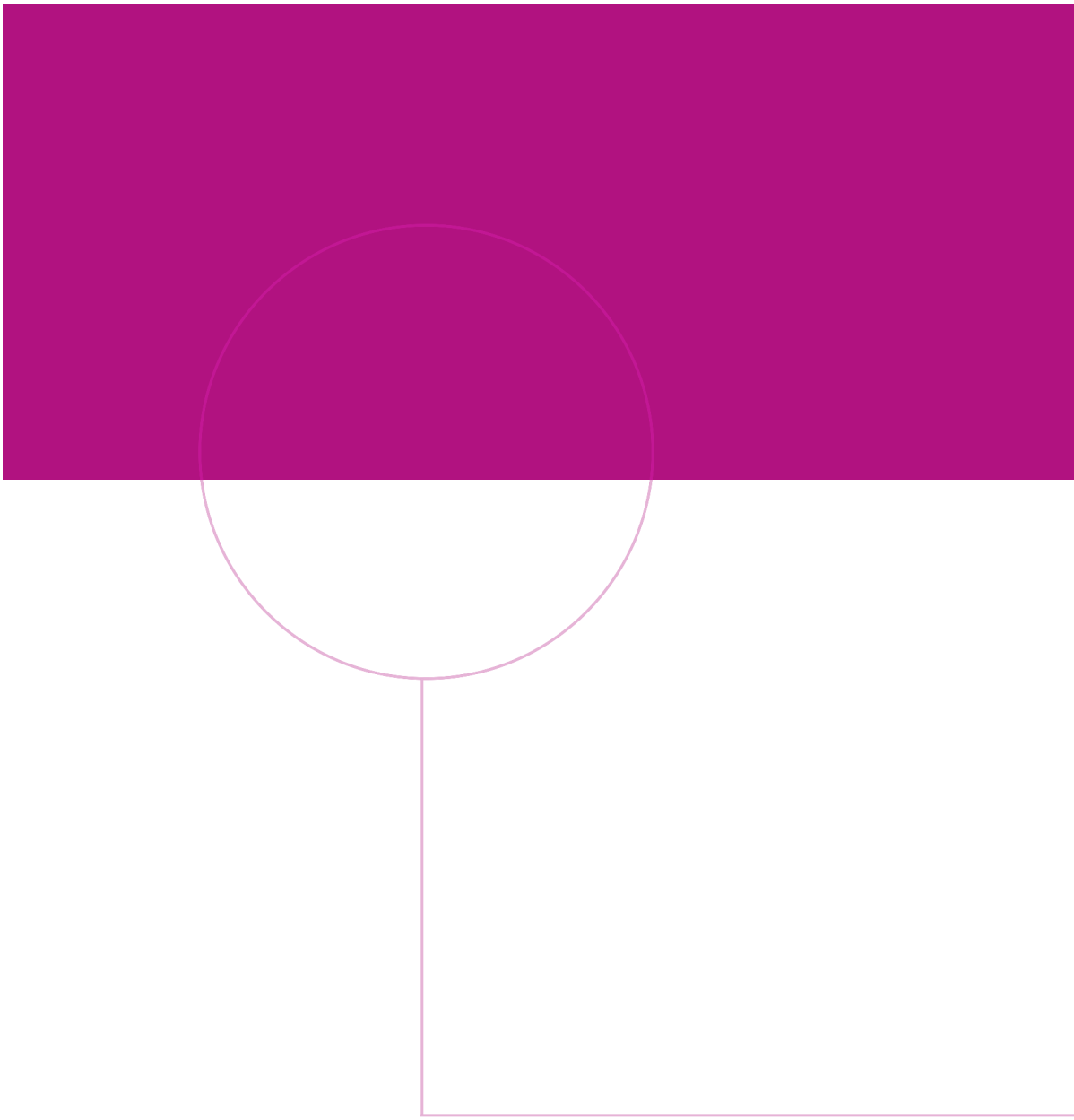
B Interviewguide for the developers

Step 1: Introduction

- Can you state your name and work title?
- Can you describe what you do in your work?
- Why did you join the project??

Step 2: Feedback on system

- To what extent do you collaborate with AI researchers and the clinicians/users of the system?
- How do you interpret the EU MDR and the AI ACT, and how do you apply it in your development of the system?
- Where does the system provide explanation? Do you have any examples? (Transparency)
- What are the outcomes of the system?
- What is the accuracy?
- Does the system give information regarding any inaccuracy in the result? (Explainability)
- XAI methods: Which one is implemented? Why?
- How do you consider the the tradeoff between the performance and the explainability? The algorithm's complexity against the possibility of explainability and transparency?
- Every Approach/implementation methods has its weaknesses, which one have you discovered? F.eks : Ignoring bias terms in the deep learning models could provide the wrong input feature attributions.
- Which security threats are you most exposed to and/or what do you do to prevent this?
- Do you teach the users how the system works? If so, how does this take place?
- What dataset do you use? How do you prevent bias?
- AI systems should not only solve pattern recognition problems but also provide causal models of the world that support explanation and understanding, how do you take this into account?
- Does the system provide options for help if it is not understandable?
- How do you try to achieve transparency, explainability, prevent bias, privacy and transferability?



NTNU

Kunnskap for en bedre verden