BACHELOR

Assessing Quality, Trust, and Behavioral Intention of XAI Methods in Healthcare

Vogten, Isa

Award date:
2024

Assessing Quality, Trust, and Behavioral Intention of XAI Methods in Healthcare

Isa Vogten

Eindhoven University of Technology

JBP000

dr. D. Sent

June 17, 2024

Abstract

With the rapid advancement of Artificial Intelligence, numerous practical applications of AI systems with high performance have emerged in the medical field, underscoring the necessity for effective collaboration between physicians and AI. Explainable AI (XAI) has been proposed as a method to make AI systems more interpretable, potentially enhancing the human-AI decision-making process. However, for these systems to be effectively implemented, they must first be accepted by physicians. This paper investigates the impact of four different XAI methods (Feature Importance, Decision Trees, Counterfactuals and Similar Cases) on the perceived quality, trust, and behavioral intention of physicians regarding AI-generated medical diagnoses. In this research a structured questionnaire showed these different XAI methods, each one explaining a diagnosis made by an XGBoost model. This questionnaire revealed that the perceived quality and trust in the XAI system significantly increase physicians' behavioral intention to use these systems. This finding underscores the importance of designing XAI systems that enhance trust and perceived quality to facilitate their integration into healthcare settings. Furthermore, AI decisions with explanations using Feature Importance and Decision Trees were perceived as higher in quality than the AI decisions without explanations. However, no significant differences in trust or behavioral intention were found between the XAI methods. This result suggests that further research is needed to determine how XAI can be effectively implemented in the medical sector to optimize its acceptance and usage.

Assessing Quality, Trust, and Behavioral Intention of XAI Methods in Healthcare

1 Introduction

The healthcare industry is changing, as costs are rising and there is a lack of health-care experts. This shows the need for new, information technology-based solutions (Yousef Shaheen, 2021). In recent years, the interest for Artificial Intelligence in the medical field has grown. Recent research highlights a great potential of using AI in the medical field (Argho et al., 2024, Kumar and Gandhi, 2018).

Reddy et al. (2019) have identified four areas where AI-enabled healthcare delivery is likely to have the most influence: healthcare administration; clinical decision support; patient monitoring; and healthcare interventions. Especially in the last three areas a lot of risk is involved. For these high-risk applications, the European Union has developed regulations for AI (Apostle et al., 2024). These include the need for human oversight: 'High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use'. Human oversight can be achieved using multiple mechanisms. Firstly, the human-in-the-loop mechanism involves a human's ability to intervene in every decision cycle of the system. Human-on-the-loop pertains to a human's capacity to intervene during the design cycle and to monitor the system. Secondly, human-in-command refers to the capability to oversee the overall activity of the AI system and determine when and how the system should be used. AI-enabled healthcare needs to have one of these mechanisms in practice. Therefore, human-AI collaboration is essential for the implementation of AI systems in the medical sector.

It is quite rational to combine human intelligence with Artificial Intelligence. Dellermann et al. (2019) discuss the idea of Hybrid Intelligence: combining the complementary capabilities of humans and AI to augment each other. The human capabilities include flexibility, creativity, empathy, the ability to adapt to different situations and common sense. These capabilities are useful when a problem deviates from the problems an AI was trained on. However, Artificial Intelligence wins from humans at processing large amounts of data and finding patterns. AI

reliably processes information, whereas human information processing is influenced by numerous factors, such as their mood, level of attention, and environmental stimuli. Moreover, humans are subject to a variety of cognitive biases (Daniel Kahneman, 2011). These complementary capabilities of humans and AI have resulted in a wide range of applications. One area of application is the medical sector, where AI can assist medical professionals in clinical decision-making. Previous studies have shown that joint human-AI medical decision-making outperformed the two decision-making systems on their own. Reverberi et al. (2022) researched the performance of endoscopists and AI in a colonoscopy diagnosis. They found that joint human-AI diagnosis outperformed both parties on their own. They claim that this is facilitated by the ability of endoscopists to distinguish the good and bad AI advice. A research by Steiner et al. (2018) lets pathologists detect breast cancer from images. Some pathologists were assisted by AI that identified and outlined regions with high likelihood of containing tumors. These AI-assisted pathologists outperformed both parties alone. They explain that this could be caused by the pathologist's ability to contextualize the therapeutic implications of false positives and false negatives. Thus maximizing the sensitivity and specificity of the human-AI diagnoses. These examples of joint human-AI collaboration highlight the potential for human-AI collaboration in the field of healthcare.

However, in order for human and AI to collaborate in medical decision-making, the AI-system needs to be accepted by medical professionals. These medical professionals often do not have expert knowledge on AI, but need to be able to interact with the AI system for it to be effective. In order to promote trust in and acceptance of AI, Explainable AI methods have been developed to obtain human-interpretable models (Dhiman et al., 2023). A study by Baroni et al. (2022) has shown that the perceived quality of XAI and the trust in XAI have a positive impact on the intention to use the system.

While existing research highlights the promising potential of AI applications in the medical domain, there remains a need to improve the quality of AI output and the trust among users, particularly medical professionals. Despite the recognized importance of Explainable AI

(XAI) in fostering trust and behavioral intention of AI systems in healthcare, there is a lack of clarity regarding which specific XAI methods can effectively promote trust and behavioral intention in medical diagnosis systems. Therefore, this study aims to fill this gap by identifying and evaluating XAI methods suitable for enhancing quality, trust and behavioral intention in medical diagnosis systems, to facilitate the effective implementation of AI technologies in healthcare settings. It does so by answering three research questions. Firstly, this research examines what XAI methods for tabular data enhance trust, quality and behavioral intention for AI systems among medical professionals most, in the context of medical diagnosis. Secondly, it researches how the perceived output quality of XAI systems and trust in XAI systems affect the behavioral intention of such systems among medical professionals. Lastly, it examines how the initial attitude towards AI and knowledge on AI affect the behavioral intention of using an XAI System. Answering these research questions will provide insights into the factors that could enhance behavioral intention for using XAI in healthcare. Additionally, they will reveal which specific XAI methods most effectively improve behavioral intention, which is crucial given the numerous available XAI methods.

2 Preliminaries

This section provides an overview of the foundational concepts relevant to the acceptance, trust and evaluation of XAI systems in healthcare.

2.1 AI in healthcare

Kumar and Gandhi (2018) highlight the potential of using information technology (including AI) in the medical field. They describe a novel three-tier Internet of Things architecture that diagnoses heart diseases in patients. In this way, huge volumes of tabular health data can be processed quickly to result in a diagnostic recommendation.

The application of AI in the field of healthcare goes much further than processing tabular data. Artificial Intelligence can also be applied to classify medical images. For example, Argho et al. (2024) have developed a Convolutional Neural Network that classifies lung X-ray images into four different categories, one of which is COVID-19. This AI model achieved an accuracy of

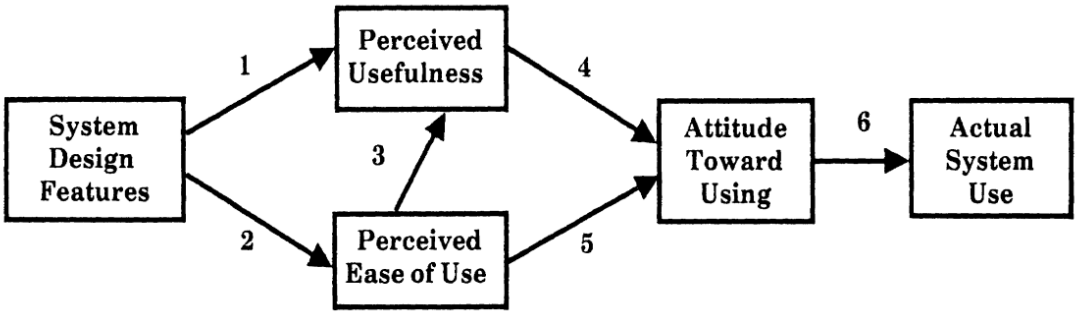
96.12% in predicting lung diseases. Implementing this neural network could replace the time-consuming RT-PCR testing system, which is a big step towards making healthcare more efficient. These developments highlight the potential of Artificial Intelligence in the healthcare domain.

2.2 Technology Acceptance

In order to be used effectively, AI systems need to be accepted by medical professionals first. The Technology Acceptance Model (TAM) developed by Davis (1987) was created to model the factors that determine user acceptance of technology. Davis provided validated measurement tools of the constructs 'perceived usefulness' and 'perceived ease of use'. These constructs were hypothesized to be determinants to the acceptance of technology. Perceived usefulness is defined as: 'the degree to which an individual believes that using a particular system would enhance his or her job performance'. The perceived ease of use is defined as 'the degree to which an individual believes that using a particular system would be free of physical and mental effort'. The exact model proposed, can be seen in Figure 1a.

Based on the original TAM model, Baroni et al. (2022) have developed a TAM model for human-in-the-loop AI applications (AI-TAM). Next to modelling perceived usefulness and ease of use, they also model two other variables related to AI. Namely, they model AI Output Trust and perceived AI Output Quality. They make a distinction between Behavioral Intention and Collaborative Intention. Behavioral Intention is the perceived probability that the user will use the system. The collaborative intention is added to also measure the willingness to participate in human-in-the-loop-AI, thus the willingness to contribute to the continuous improvement of the AI results. They evaluated this model using a real-world application called "BumpOut". Users went through a car damage application for insurance reasons. AI assisted in analysing the severity of the damage, using an image of the damage. Users could confirm or deny the suggestion of the AI. After this task, users were given a questionnaire to measure the constructs in the AI-TAM model. Their AI-TAM model proposed, with the regression loading variables is shown in Figure 1b. Note that AI Output Trust and AI Output Quality influence the Behavioral Intention, but that no

(a) *Technology Acceptance Model (TAM) by Davis (1987)*



(b) *AI Technology Acceptance Model (AI-TAM) by Baroni et al. (2022)*

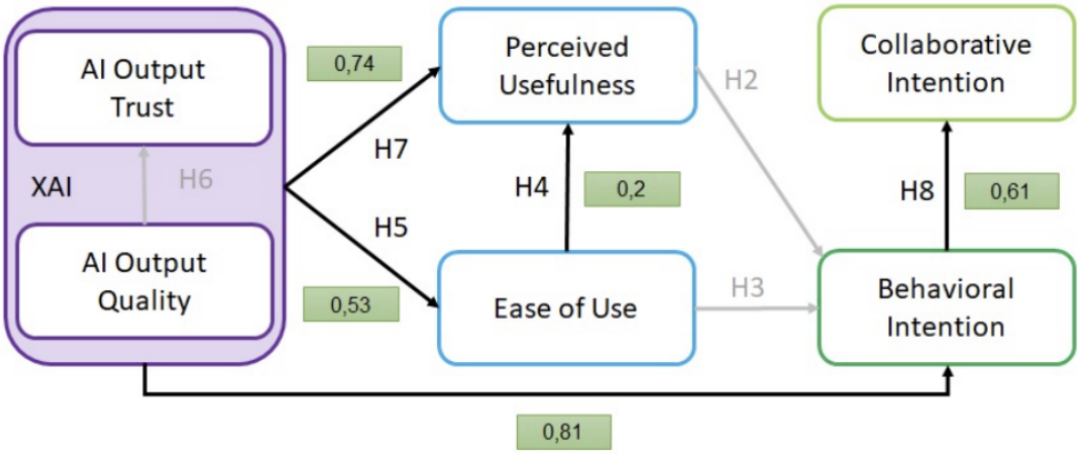


Figure 1

Technology Acceptance Models

evidence was found that Perceived Usefulness and Ease of Use influence the Behavioral Intention. Furthermore, Behavioral Intention influenced Collaborative Intention positively. The researchers found that AI Output Quality and AI Output Trust could not be completely distinguished from each other, because of a correlation of 1. However, their AI-TAM model reveals a strong positive effect of AI Output Trust and AI Output Quality on the intention to use AI.

2.3 Trust in AI

From the AI-TAM model, it can be understood that trust in the output of the AI determines the behavioral intention for the AI system. Mayer et al. (1995) define trust as 'the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will

perform a particular action important to the trustor, irrespective of the ability to monitor or control that other part'. Choung et al. (2023) have also investigated the role of trust in the use of AI. They showed a significant effect of trust on the intention to use AI voice-assistants. Asan et al. (2020) claim that: 'Currently, a lack of trust in the AI systems is a significant drawback in the adoption of this technology in healthcare.' So, in order to be implemented in healthcare systems, AI systems need to be trusted by users. Research has shown that AI is not widely trusted in the medical domain. Jungmann et al. (2021) conducted a survey to investigate the attitudes towards AI and its future impact on radiology. The survey was employed amongst radiologists, IT specialists and industry representatives. The results showed that participants were convinced that AI would make radiology more efficient. However, participants had a low level of trust in the results of AI solutions. Weber et al. (2024) found another factor that influences trust in AI. They conducted a qualitative and a quantitative study to investigate attitudes and opinions towards adopting AI-enabled healthcare technologies in their work. They analyzed Reddit threads of physicians about AI and conducted an online survey. Their work has shown that the fear of being replaced by an AI and skepticism towards AI played major roles in conversations by medical professionals. Weber et al. (2024) have also found that: 'the intention [of medical professionals] to use AI technologies increases with increasing knowledge about AI and this effect is moderated by the fear of being replaced by AI.' This shows that the initial attitude towards AI and knowledge on AI can have implications for the amount of trust medical experts have in these AI systems.

2.4 Explainable AI in Healthcare

Dhiman et al. (2023) state that the lack of trust in AI models in the healthcare domain is caused by the black-box nature of AI. This black-box nature is caused by the complexity of the models, which makes it impossible for humans to interpret them.

Rai (2020) describes a trade-off between prediction and explanation; simple machine learning algorithms can produce interpretable models, but provide less accurate predictions than the difficult-to-interpret deep learning models. To summarize, improving AI performance often reduces interpretability, while enhancing interpretability can limit performance.

Rai discusses Explainable Artificial Intelligence (XAI) as a method to explain the logic behind the AI system and to help integrate XAI in the medical domain. The main goal of Explainable AI is to obtain human-interpretable models. This may have many advantages, which include combating negative consequences of automated decision-making, helping individuals to make more informed choices and integrating algorithms with human values (Ali et al., 2023).

A literature study by Wen Loh et al. (2022) describes two different approaches to XAI: global explanations and local explanations. Global explanations try to make the whole model transparent. Local explanations explain how a specific input influences one single model prediction.

An experiment by Panigutti et al. (2022) showed a diagnosis of an AI system with an explanation to healthcare providers. This local explanation highlighted the conditions of the patient the AI system deemed most important for the patients diagnosis. Furthermore, a textual explanation was given for each important condition. This research has shown that providing these explanations to healthcare providers increases the influence of the AI suggestion. They did not find an increase of intention to use by providing the explanations to AI suggestions. However, they state that this could be due to an unsatisfactory explanation.

Wysocki et al. (2022) investigated the role of Machine Learning explanations in the field of healthcare. Despite a positive attitude towards AI explanations of healthcare professionals, they found negative effects of XAI. These effects were associated with confirmation bias, accentuated model over-reliance and an increased effort to interact with the model. If XAI will be implemented in healthcare, these potential issues need to be taken into account.

Wen Loh et al. (2022) have identified the 10 most implemented XAI methods in research on the healthcare industry. These methods can be categorized by the type of data they process. The first category processes images. This category entails GradCAM, Layer-wise Relevance Propagation and other heatmap and saliency map generation methods. Another category processes tabular data. This category exists of SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016), Fuzzy Classifiers (Piotr Prokopowicz et al., 2017), Explainable Boosting Machines

(Lou et al., 2013), Case-Based reasoning and Rule-Based systems. The last category processes natural language, which can use stylometry or NLP. These popular XAI methods for each data type are summarized in Table 1.

Table 1

Popular XAI methods per data type

Data Type	Popular XAI Methods
Images	GradCAM, Layer-wise Relevance Propagation
Tabular data	SHAP, LIME, Fuzzy Classifiers, Explainable Boosting Machines, Case-Based reasoning and Rule-Based systems
Natural language	NLP, stylometry

This research will focus on XAI for tabular data. Within this category, multiple ways of communicating explanations are possible. Firstly, SHAP and LIME both provide explanations by explaining the important features of a model or prediction. Where SHAPs explanation is for a complete dataset, LIMEs explanation is only local (Wen Loh et al., 2022). These algorithms calculate a feature importance score for every feature in the model. Das and Wiese (2023) propose to rank features in descending order, based on their importance. This helps medical practitioners with prioritizing their attention.

A second type of explanation uses rules or Decision Trees to explain AI-systems. Examples are Fuzzy classifiers (Chimatapu et al., 2018), Explainable Boosting Machines (Nori et al., 2019) and Rule-Based systems. Panigutti et al. (2020) have developed *doctor XAI*, an XAI system that provides a rule-based explanation for a classification prediction in the field of oncology.

Another method, called Case-Based reasoning, uses similar cases. It shows a similar case, for which the target variable is known, to the user as explanation for the prediction of a case for which the target variable is unknown. Caro-Martinez et al. (2019) have developed a method to

select similar cases to the case that has been predicted by an AI model using link prediction techniques.

Lastly, Counterfactuals show how the input would need to change to get a specific output (Baron, 2023). Unlike sensitivity analysis, which examines the impact of varying each input variable on the output, Counterfactuals identify precise changes needed for a particular outcome. Often, the data of a patient is adjusted to create a fictional what-if situation. All these types of communication of XAI outcomes are summarized in Table 2. Examples of popular XAI methods for each communication method are also given.

These studies have shown the potential for XAI to enhance trust. However, with the broad scale of XAI methods, it is still unclear which methods work best.

2.5 XAI evaluation

Some difficulties arise when trying to evaluate the performance of different XAI systems. What is needed for an XAI system to be good? Hoffman et al. (2018) explore metrics for assessing the effectiveness of XAI systems, aiming to measure their quality and performance.

Especially interesting for this research are their metrics for explanation satisfaction and trust, as Baroni et al. (2022) state that it is the AI Output Quality and AI Output Trust that determines Behavioral Intention, as explained in Section 2.2. Hoffman et al. (2018) define the satisfaction of the explanations as the degree to which the users feel that they understand the AI system that is explained to them. Their paper proposes a scale to measure this construct. They have also developed a scale to measure trust in XAI.

Table 2

Ways of communicating XAI explanations for tabular data and their explanation

Type of communication	**Explanation**	**Popular Methods**
Feature Importance	Identifies which features have the highest impact on a model's predictions.	SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016)
Rules or Decision Trees	Describes which rules a model uses to predict.	Fuzzy Classifiers (Chimatapu et al., 2018, Explainable Boosting Machines (Nori et al., 2019)
Case-Based Reasoning	Shows previous similar cases and their outcome to explain the prediction of a new case.	Similar Case Explanations (Caro-Martinez et al., 2019
Counterfactuals	Give insight into how the input needs to change to get a certain output.	Causal Counterfactuals (Chou et al., 2021)

3 Methods

To answer the three research questions, a questionnaire with a within-subjects design with five factors was conducted among physicians. The five factors were an AI baseline and four presentations of different XAI methods.

3.1 Participants

Participants were gathered using a connection at Thuisarts.nl, a platform that shares reliable information about health and illness gathered from 40 different associations of physicians in the Netherlands ("Over Thuisarts.nl | Thuisarts.nl," n.d.). Next to Thuisarts.nl, LinkedIn was used to share the survey with medical professionals ("LinkedIn," n.d.). The description of the survey stated that participants needed to have a medical degree, or were medical students.

3.2 Measures

Initial Measures. Weber et al. (2024) have found that the initial attitude towards AI and knowledge on AI can have implications for the amount of trust medical experts have in AI systems. Therefore, to account for differences in prior attitude towards AI and knowledge on AI, these constructs were measured. The attitude towards AI was measured using a measurement tool by Grassini (2023). It is shown in Appendix B. A tool used by Weber et al. (2024) was used to measure the knowledge on AI. It is shown in Appendix A.

XAI-related measures. For each XAI communication method, output quality, output trust and behavioral intention were measured. Output quality was measured using an adapted version of the explanation satisfaction scale by Hoffman et al. (2018). Then, trust in the XAI was measured using an adaptive version of the scale of Hoffman et al. (2018). Last, intention to use the XAI system was measured using an adapted scale of Venkatesh et al. (2003). Not all items of before mentioned scales were used, to keep the questionnaire short. The selection can be found in Appendices C, D, E.

3.3 XAI prototypes

This section explains the approach to training the model and creating the images for each of the the XAI communication methods.

3.3.1 AI model

All prototypes were based on an XGBoost classification model aimed to diagnose patients with Chronic Kidney Disease (CKD). The XGBoost model uses gradient boosting trees to predict whether a patient has CKD or not (Chen and Guestrin, 2016). The model was trained on data in a database of Rubini et al. (2015). The Chronic Kidney Disease dataset was selected for this study because of its ease of use, requiring minimal data engineering. Additionally, this dataset has been effectively utilized in prior research for training an XGBoost model, demonstrating its suitability and reliability for this application (Raihan et al., 2023). This data contains 24 features of health measurements and an indicator of CKD. All missing values were removed, this resulted in a dataset of 230 rows. Earlier research has found that CKD can be diagnosed quite accurately using a XGBoost model and nine features (Raihan et al., 2023). Based on that research, the same nine features were chosen for model training. These features include: Blood pressure (mm/Hg), Albumin level (1-5), Bacteria present, Blood Glucose Random (mgs/dL), Blood ureum (mgs/dl), Potassium (mEq/L), Diabetes Mellitus (yes, no), Coronary Heart Disease (yes, no).

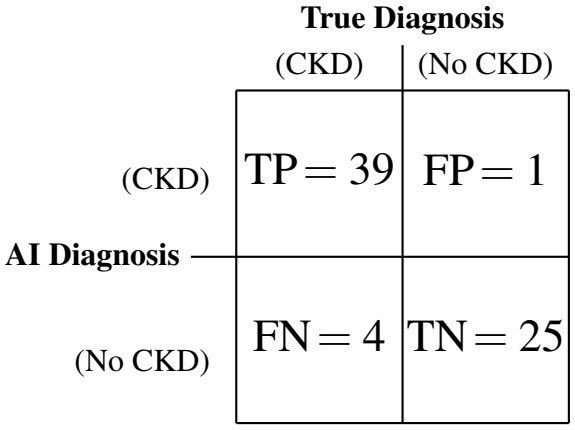
The model was trained on 70% of the data and tested on 30% of the data. The confusion matrix for the classification of the test set can be found in Figure 2. Overall, the model had an accuracy of 93%.

3.3.2 Feature Importance explanation

The explanation using Feature Importance was made using the SHAP library, specifically the waterfall plot was used (Lundberg and Lee, 2017). This method takes the nine measurements for a single patient as input and explains how these measurements influence the decision of the model for that patient. It does so by showing the effect of all measurements on the probability of getting a CKD prediction, starting with the most influential measures. The vertical line shows the final chance of getting a CKD prediction, which is 0.998 in this case. The colors and font of the graph

Figure 2

Confusion matrix for the predictions of test set from XGBoost model



were changed in order to have a similar design as other explanation methods and reduce bias based on the appearance of the presentations. The final graph can be seen in Figure 3a.

3.3.3 Rule based explanation

The explanation based on rules was provided using a Decision Tree. The XGBoost model uses a set of Decision Trees to come to a prediction. To translate this set of Decision Trees to one tree, an approach of Sagi and Rokach (2021) was used. This approach creates an interpretable Decision Tree from Gradient Boosting Decision Trees, such as XGBoost, to increase interpretability. The Decision Tree differs from the other explanations, because it is both a local and a global explanation; it can explain the overall model behavior, but also a single prediction. The Decision Tree can be found in Figure 3b.

3.3.4 Similar Case explanation

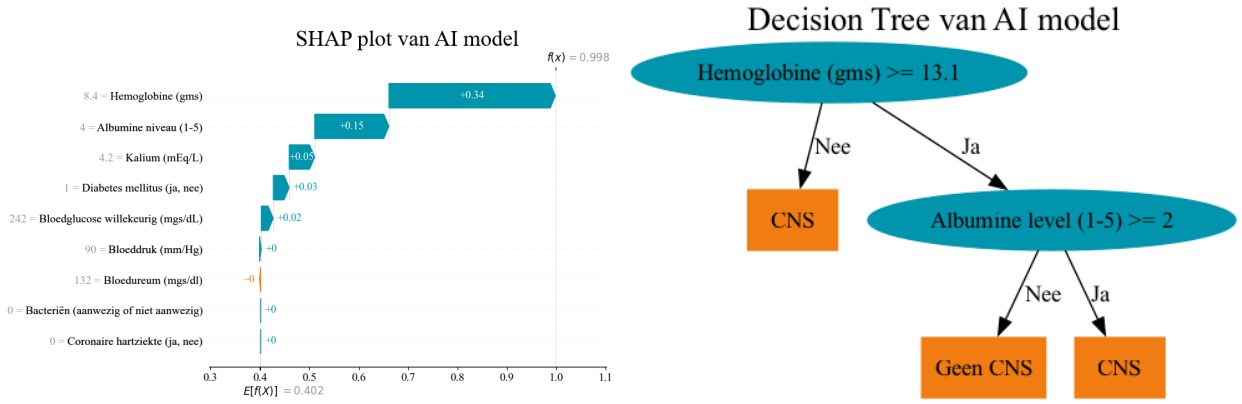
The explanation using a Similar Case was plotted manually, using the library Matplotlib (Hunter, 2007). To find similar cases, all features were normalized to ensure all features contributed equally to the distance calculations. Then, the k-nearest neighbour algorithm was used to find the closest neighbour for each data point. Then, one of these data points and its closest neighbour were plotted. This can be seen in Figure 3c.

3.3.5 Counterfactual explanation

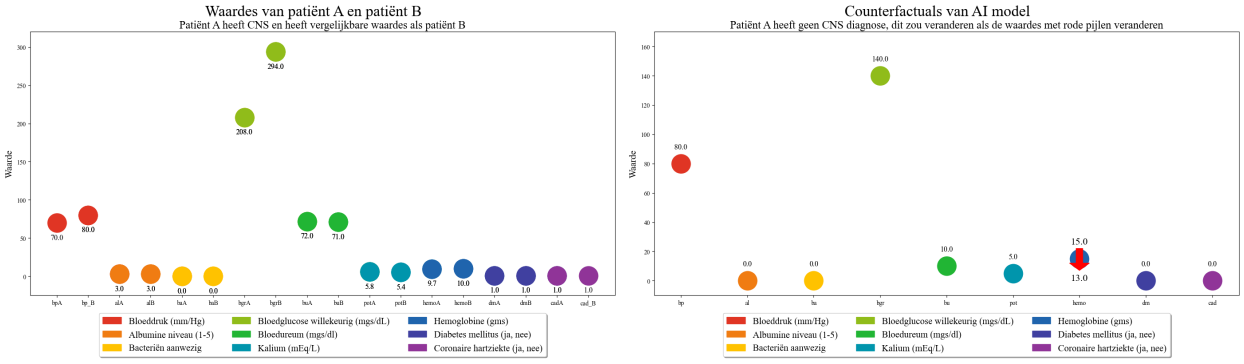
The Counterfactual explanation was manually constructed using the Decision Tree described in Section 3.3.3. A data point with a 'No CKD' prediction was chosen, and the Decision Tree rules were used to decide what changes were needed for this data point to get a CKD prediction. These changes were visualized using red arrows, this can be seen in Figure **??**.

Figure 3

Presentations of different XAI methods



(a) *Feature Importance* (b) *Decision Tree*



(c) *Similar Case*

3.4 Survey

To administer the survey, an online survey was used because this would give access to a broad group of specific people: physicians (in training) (Wright, 2006). The survey platform Qualtrics

was used to administer the survey. It is user-friendly and GDPR-compliant ("Qualtrics XM - Experience Management Software," 2024). The survey was anonymous. The survey was conducted in Dutch, to make sure participants were not hampered by language. The different parts of the survey will be discussed individually in the following sections.

Start. At the beginning of the survey, participants were given an introduction to the survey and the matter of data processing. Then, they were asked for their informed consent. Furthermore, gender, age, and occupation.

Initial Attitude towards AI. First, a questionnaire was conducted to measure feelings and attitudes towards AI, based on a scale proposed by Grassini (2023). It consisted of 3 items, which can be found in Appendix B. Then, another questionnaire was used to assess the knowledge participants had on AI, based on a scale proposed by Weber et al. (2024). It consisted of 4 items and can be found in Appendix A. Both questionnaires used a 5-point Likert-Scale to collect answers.

Baseline As a baseline, participants were shown the outcome of the AI system without explanation. For this outcome, the satisfaction, trust and behavioral intention were measured using a five-point Likert-Scale.

Examples of Medical Diagnosis AI Systems. Then, the XAI communication methods in Table 2 were explained and shown one by one in random order. After showing the XAI communication methods, for each method, three constructs were measured. Firstly, the explanation satisfaction with the XAI system was measured using an five-item scale. It used a five-point Likert-Scale and it can be found in Appendix C. Secondly, trust in the XAI system has been measured using an eight-item trust scale. This scale can be found in Appendix D. Lastly, the behavioral intention to use the system has been measured using the question in Appendix E.

End. After the survey, participants were thanked for their time.

3.5 Data Analysis

This section describes the method of data analysis. It starts with the data preprocessing and then continues to the statistical tests that have been used.

3.5.1 Preprocessing of results

Means were taken of the Likert-scale items for attitude, knowledge, satisfaction, trust and behavioral intention. If there was one missing item for a construct that seemed random, the mean was calculated using the other items of that construct.

Not all responses were complete. Only responses that answered the questions for the baseline and at least one of the four XAI methods were taken into account.

3.5.2 Statistical inference

Descriptive statistics were used to summarize participant demographics, initial attitude towards AI, initial knowledge on AI, explanation satisfaction, trust in AI and behavioral intention. Inferential statistical tests were performed to examine the differences in explanation satisfaction, trust and behavioral intention between different XAI methods. All assumptions of statistical tests were checked beforehand.

To determine whether there were differences in quality or trust between the five conditions, two one-way ANOVAs were performed. This is possible because parametric analysis of the mean of multiple Likert-Scale items can be justified by the Central Limit Theorem (University of St. Andrews, n.d.). To determine whether there were differences in behavioral intention, a Kruskal-Wallis H Test was performed, because behavioral intention was an ordinal variable. If significant effects were found, a post-hoc test was done to determine which conditions differed from each other.

To investigate the factors that influenced behavioral intention, an ordered regression analysis was performed to estimate the relationships between a set of independent variables and behavioral intention. The independent variables in the set were age, gender, initial attitude towards AI, initial knowledge on AI, perceived output quality and trust in AI. The coefficients of the parameters and their significance were used to answer these research questions.

4 Results

This section will describe the descriptive statistics first. Afterwards, the results of the statistical tests that were performed are shared.

4.1 Descriptive Statistics

35 participants were recruited over a period of 14 days. The analysis only considered the results from participants who completed all constructs for the baseline and at least one of the XAI methods. This resulted in a dataset of 19 records. Table 3 shows the distribution of occupation and gender within the sample. Figure 4 shows the age of these participants. Most participants were male and between 40 and 49 years of age.

Description	**Distribution**	
Occupation	Physician: 89%	Physician in training: 11%
Gender	Male: 58%	Female: 42%

Table 3

Demographic information

Table 4, 5, 6, 7 show the descriptive statistics of the attitude towards AI, knowledge on AI, and quality, trust and behavioral intention of baseline and the different XAI methods.

Table 4

Descriptive statistics of initial attitude towards AI and knowledge on AI (1-5)

	attitude	knowledge
count	19.00	19.00
mean	4.09	3.69
std	0.55	0.73
min	3.00	2.33
median	4.33	3.50
max	5.00	5.00

Table 5

Descriptive statistics of perceived quality with the baseline, Feature Importance explanation, Decision Tree explanation, Similar Case explanation and Counterfactual explanation

	quality baseline	quality FI	quality DT	quality SC	quality CF
count	19	18	16	16	17
mean	2.09	2.97	2.86	2.49	2.27
std	0.78	0.99	1.01	0.95	0.73
min	1.00	1.00	1.00	1.00	1.00
median	2.00	3.10	2.80	2.30	2.20
max	3.80	5.00	5.00	4.40	3.80

Table 6

Descriptive statistics of trust in the baseline, Feature Importance explanation, Decision Tree explanation, Similar Case explanation and Counterfactual explanation

	trust baseline	trust FI	trust DT	trust SC	trust CF
count	19	18	16	16	17
mean	2.23	2.72	2.83	2.58	2.37
std	0.79	0.80	0.84	0.68	0.56
min	1.00	1.00	1.25	1.50	1.50
median	2.00	2.88	2.88	2.50	2.25
max	3.50	4.00	4.00	3.75	3.50

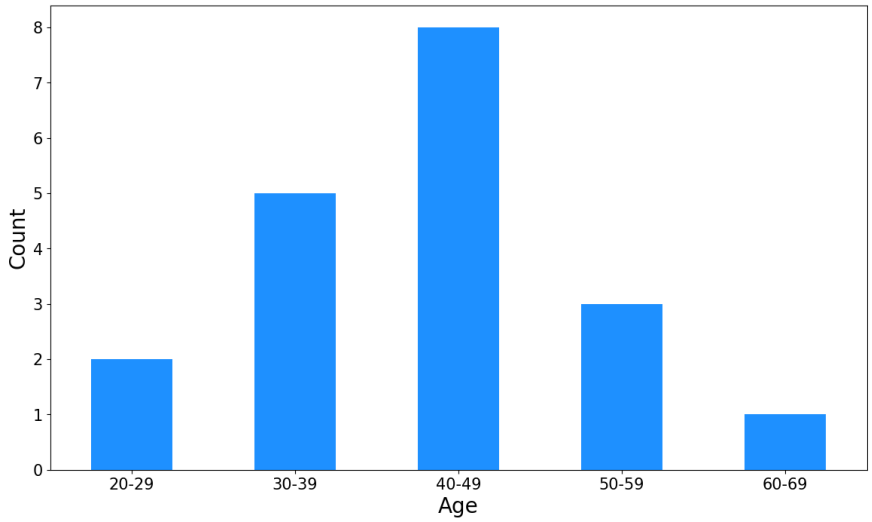
Table 7

Descriptive statistics of behavioral intention for the baseline, Feature Importance explanation, Decision Tree explanation, Similar Case explanation and Counterfactual explanation

	bi base	bi FI	bi DT	bi SC	bi CF
count	19	18	16	16	17
mean	2.63	3.00	2.94	2.69	2.35
std	1.21	1.19	1.24	1.01	0.93
min	1.00	1.00	1.00	1.00	1.00
median	3.00	3.00	3.00	3.00	2.00
max	5.00	5.00	5.00	4.00	4.00

Figure 4

Bar chart age of participants



4.2 Difference in trust, quality and behavioral intention between the 5 conditions

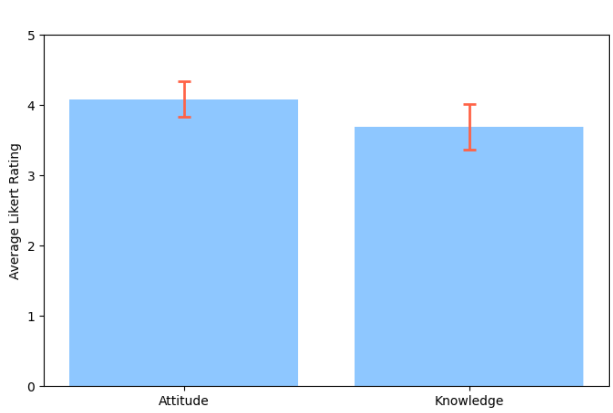
The mean initial measures, trust, quality and behavioral intention in all 5 conditions is shown in Figure 5.

To conclude whether there is a significant difference in quality and trust between the five different presentations of the AI output, two one-way ANOVA's were performed. Before the test, the normality and homogeneity of variance were checked using a Shapiro-Wilk test and a Levene test. All observations were independent, were normally distributed, and had homogeneity of variance. No significant difference was found in trust between the five presentations of AI output (p = 0.108). Therefore, no post-hoc analysis was done on the trust measure. However, a significant difference in quality was found (p = .020). A post-hoc t-test analysis shows that the quality of the Decision Tree XAI method (M=2.814, SD = 1.027) was significantly higher than the quality of the baseline (M=1.900, SD = 0.722), t(33) = -2.545, p = .016. Moreover, the quality of the Feature Importance XAI method (M=3.029, SD = 1.064) was significantly higher than the quality of the baseline (M=1.900, SD = 0.722), t(35) = -2.987, p = .005. For this post-hoc analysis test, t-tests with a Bonferroni correction for multiple comparisons was used.

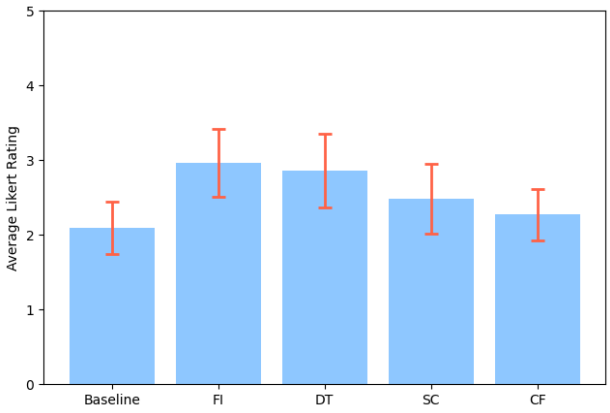
To conclude whether there is a significant difference in behavioral intention between the

Figure 5

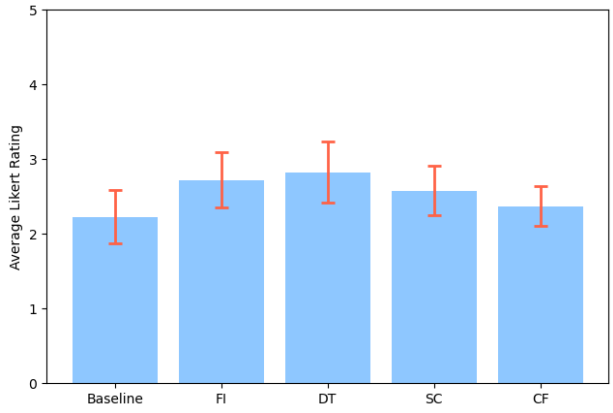
Bar plot of mean trust per XAI method



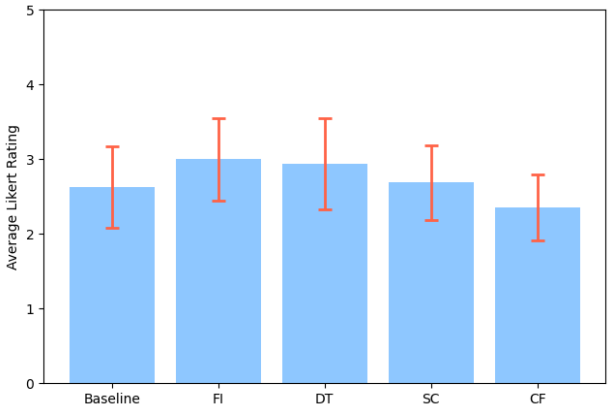
(a) *Bar plot of initial measurements*



(b) *Bar plot of mean quality per XAI method*



(c) *Bar plot of mean trust per XAI method*



(d) *Bar plot of mean behavioral intention per XAI method*

five different presentations of the AI output, a Kruskal-Wallis H Test was performed. This test is non-parametric, thus can be used to make inferences on the behavioral intention, which is an ordinal variable. No significant difference in Behavioral Intention was found (p = .338). The results of the two ANOVA's and the Kruskal-Wallis H test are summarized in Table 8.

Table 8

Tests for differences in trust, quality and behavioral intention

Test	Test Statistic	p-value
one-way ANOVA quality	3.104	0.020
one-way ANOVA trust	1.964	0.108
Kruskal-Wallis H BI	3.370	0.338

4.3 Factors influencing Behavioral Intention of XAI methods

To investigate which factors influence the behavioral intention, an ordinal regression was performed with behavioral intention as the target variable. The predictor variables were age, attitude, knowledge, gender, quality and trust. The coefficients, z-values and p-values of the parameters can be seen in Table 9

Because of a high correlation between quality and trust ($\rho = 0.789$), these variables were added and seen as one variable to prevent multicollinearity.

One can notice that no significant effect for age was found. However, attitude, knowledge, gender and quality and trust had a significant effect on the behavioral intention of physicians to use XAI systems. Male physicians had a lower behavioral intention than female physicians. Furthermore, knowledge and quality and trust had a positive effect on the behavioral intention of physicians. However, a positive attitude towards AI resulted in a lower behavioral intention.

Table 9

Parameters of ordered regression to predict behavioral intention

	Coefficient	**Std. Dev.**	**Z-value**	**p-value**
age = 20-29	0.0691	0.820	0.084	0.933
age = 30-39	-0.3066	0.658	-0.466	0.641
age = 40-49	-0.8971	0.614	-1.461	0.144
age = 50-59	0.0871	0.635	0.137	0.891
attitude	-0.5843	0.290	-2.015	0.044
knowledge	1.0940	0.308	3.555	< 0.001
gender = Male	-0.7859	0.387	-2.030	0.042
quality + trust	0.9947	0.134	7.431	< 0.001

5 Discussion

Table 8 shows the results of the two one-way ANOVA tests and the Kruskal-Wallis H Test that were performed to compare the five conditions. These tests examined the differences in mean quality, trust, and behavioral intention among five different presentations of AI outputs: the baseline AI, the Feature Importance explanation, the rule-based explanation, the Similar Case explanation, and the Counterfactual explanation. It shows that the one-way ANOVA on quality of the XAI methods resulted in a p-value < 0.05. This means that there was a difference in quality between the five AI output presentations. A post-hoc test shows that the Feature Importance and Decision Tree XAI methods had a significantly higher mean quality than the baseline. It cannot be concluded that there was a difference in trust or behavioral intention between the five AI output presentations.

Table 9 provides insights from an ordinal regression analysis on behavioral intention. It shows that the age of physicians did not affect behavioral intention. Additionally,the quality of and trust in the XAI system were found to have a positive and significant impact on behavioral intention. The initial attitude towards AI and knowledge on AI also had a significant effect on behavioral intention, the initial knowledge on AI had a positive effect. However, the initial attitude towards AI had a negative effect. Lastly, male physicians had a significantly lower behavioral intention to use XAI methods, compared to female physicians.

These positive effects of quality and trust on behavioral intention aligns with the AI-TAM framework proposed by Baroni et al. (2022), which claims that the XAI quality and XAI trust positively impact the behavioral intention. Regarding other factors that influence the behavioral intention, the findings partly align with the findings of Weber et al. (2024). This research has found that knowledge on AI improved the behavioral intention of medical professionals, which corresponds to the findings of Weber et al. (2024). However, they also found that a positive attitude towards AI had a positive impact on the behavioral intention of medical professionals, which contradicts the findings of this research.

The results underscore the importance of perceived trust and quality in XAI systems

within the healthcare sector. These factors are crucial determinants of physicians' intentions to utilize these systems. Consequently, designers and developers of XAI systems should prioritize enhancing these attributes to foster greater acceptance and usage among healthcare professionals. The results underscore the importance of information exchange on AI in the field of healthcare, as knowledge seems to play a role in the acceptance of AI systems in healthcare.

Surprisingly, the initial attitude towards AI negatively impacted the behavioral intention for AI systems in healthcare. A possible cause is the nature of the attitude questionnaire. This questionnaire was not solely applied to AI in healthcare, but also to participants' opinion on AI in general. Because of the risks that are involved in applying AI in healthcare, compared to other sectors, participants could differ in their attitude towards AI in healthcare and AI in general. Future studies could take this difference into account by making the AI attitude scale more applicable to the healthcare sector.

Some limitations apply to this research. Firstly, it is important to recognize that the behavioral intention was measured. The behavioral intention measured may not fully translate to actual usage in practical settings, indicating a need for further longitudinal studies to assess actual behavior. Secondly, there might be a selection bias in the participant pool, caused by a difference in initial interest in AI. Physicians with a high interest in AI could be more inclined to start and pursue with the questionnaire. This is also reflected in a rather high initial knowledge and a rather positive initial attitude towards AI, which is shown in Table 4. Lastly, the questionnaire had a rather low sample size, because only few participants completed the questionnaire.

Further research with a bigger sample size should be done to explore the possibilities of XAI in the medical sector and investigate which XAI methods promote perceived quality, trust and behavioral intention most.

6 Conclusion

This research investigated the quality, trust and behavioral intention of physicians for four different XAI methods supporting medical diagnosis and explored underlying factors that influenced the behavioral intention.

Results show that there was a difference in perceived quality amongst physicians between an AI decision without explanations and two XAI methods: Feature Importance and Decision Trees both promoted the perceived quality. No significant difference was found in trust or behavioral intention.

Multiple factors were found to contribute to the behavioral intention of physicians. Firstly, the initial knowledge on AI positively impacted the behavioral intention, just like the perceived quality and trust in the XAI system. Initial attitude towards AI had a negative impact on the behavioral intention.

This research contributes to the understanding of how AI can be effectively implemented in medical diagnosis, by revealing different factors that determine the acceptance of XAI systems by physicians and identifying differences in perceived quality between different XAI methods. However, research with a broader and more diverse sample should be performed to validate these findings, also focusing on the actual deployment of XAI systems in a clinical setting.

References

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, *99*. https://doi.org/10.1016/j.inffus.2023.101805

Apostle, J., Schröder, C., Schaedler, S., & Kawkab, R. (2024). Artificial Intelligence Act.

Argho, A. G., Maswood, M. M. S., Mahmood, M. I., & Mondol, N. (2024). EfficientCovNet: A CNN-based approach to detect various pulmonary diseases including COVID-19 using modified EfficientNet. *Intelligent Systems with Applications*, *21*, 200315. https://doi.org/10.1016/j.iswa.2023.200315

Asan, O., Bayrak, A. E., & Choudhury, A. (2020, June). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. https://doi.org/10.2196/15154

Baron, S. (2023). Explainable AI and Causal Understanding: Counterfactual Approaches Considered. *Minds and Machines*, *33*(2), 347–377. https://doi.org/10.1007/s11023-023-09637-x

Baroni, I., Re Calegari, G., Scandolari, D., & Celino, I. (2022). AI-TAM: a model to investigate user acceptance and collaborative intention inhuman-in-the-loop AI applications. *Human Computation*, *9*(1), 1–21. https://doi.org/10.15346/hc.v9i1.134

Caro-Martinez, M., Recio-Garcia, J. A., & Jimenez-Diaz, G. (2019). An Algorithm Independent Case-Based Explanation Approach for Recommender Systems Using Interaction Graphs. In C. Bach Kerstin Marling (Ed.), *Case-based reasoning research and development* (pp. 17–32). Springer International Publishing.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785

Chimatapu, R., Hagras, H., Starkey, A., & Owusu, G. (2018). Explainable AI and Fuzzy Logic Systems. In D. Fagan, C. Martín-Vide, M. O'Neill, & M. A. Vega-Rodríguez (Eds.), *Theory and practice of natural computing* (pp. 3–20). Springer International Publishing.

Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2021). Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications.

Choung, H., David, P., & Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human-Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Daniel Kahneman. (2011). *Thinking Fast and Slow*. Penguin Books.

Das, P. P., & Wiese, L. (2023). Explainability Based on Feature Importance for Better Comprehension of Machine Learning in Healthcare. *New Trends in Database and Information Systems*, 324–335.

Davis, F. D. (1987). User acceptance of information systems: the technology acceptance model (TAM).

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business and Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Dhiman, P., Bonkra, A., Kaur, A., Gulzar, Y., Hamid, Y., Mir, M. S., Soomro, A. B., & Elwasila, O. (2023, October). Healthcare Trust Evolution with Explainable Artificial Intelligence: Bibliometric Analysis. https://doi.org/10.3390/info14100541

Grassini, S. (2023). Development and validation of the AI attitude scale (AIAS-4): a brief measure of general attitude toward artificial intelligence. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1191628

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jungmann, F., Jorg, T., Hahn, F., Pinto dos Santos, D., Jungmann, S. M., Düber, C., Mildenberger, P., & Kloeckner, R. (2021). Attitudes Toward Artificial Intelligence Among

Radiologists, IT Specialists, and Industry. *Academic Radiology*, *28*(6), 834–840. https://doi.org/10.1016/j.acra.2020.04.011

Kumar, P. M., & Gandhi, U. D. (2018). A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. *Computers and Electrical Engineering*, *65*, 222–235. https://doi.org/10.1016/j.compeleceng.2017.09.001

LinkedIn. (n.d.). https://linkedin.com/

Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 623–631. https://doi.org/10.1145/2487575.2487579

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. http://arxiv.org/abs/1705.07874

Mayer, R. C., Davis, J. H., & David Schoorman, F. (1995). *An Integrative Model of Organizational Trust* (tech. rep. No. 3). https://www.jstor.org/stable/258792?seq=1&cid=pdf-

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability.

Over Thuisarts.nl | Thuisarts.nl. (n.d.). https://www.thuisarts.nl/over-thuisarts

Panigutti, C., Beretta, A., Giannotti, F., & Pedreschi, D. (2022). Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3502104

Panigutti, C., Perotti, A., & Pedreschi, D. (2020). Doctor XAI An ontology-based approach to black-box sequential data classification explanations. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 629–639. https://doi.org/10.1145/3351095.3372855

Piotr Prokopowicz, Jacek Czerniak, Dariusz Mikołajewski, Łukasz Apiecionek, & Dominik Ślzak. (2017). *Theory and Applications of Ordered Fuzzy Numbers* (P. Prokopowicz, J. Czerniak, D. Mikołajewski, Ł. Apiecionek, & D. Ślzak, Eds.; Vol. 356). Springer International Publishing. https://doi.org/10.1007/978-3-319-59614-3

Qualtrics XM - Experience Management Software. (2024, May). https://www.qualtrics.com/

Rai, A. (2020, January). Explainable AI: from black box to glass box. https://doi.org/10.1007/s11747-019-00710-5

Raihan, M. J., Khan, M. A. M., Kee, S. H., & Nahid, A. A. (2023). Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-33525-0

Reddy, S., Fox, J., & Purohit, M. P. (2019, January). Artificial intelligence-enabled healthcare delivery. https://doi.org/10.1177/0141076818815510

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., Antonelli, G., Awadie, H., Bernhofer, S., Carballal, S., Dinis-Ribeiro, M., Fernández-Clotett, A., Esparrach, G. F., Gralnek, I., Higasa, Y., Hirabayashi, T., Hirai, T., Iwatate, M., Kawano, M., Mader, M., … Cherubini, A. (2022). Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-18751-2

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease.

Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522–542. https://doi.org/10.1016/j.ins.2021.05.055

Steiner, D. F., Macdonald, R., Liu, Y., Truszkowski, P., Hipp, J. D., Gammage, C., Thng, F., Peng, L., & Stumpe, M. C. (2018). *Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer* (tech. rep.). www.ajsp.com.

University of St. Andrews. (n.d.). *ANALYSING LIKERT SCALE/TYPE DATA* (tech. rep.). https://www.st-andrews.ac.uk/media/ceed/students/mathssupport/Likert.pdf

Venkatesh, V., Morris, M. G., Davis, G. B., Davis, F. D., Smith, R. H., & Walton, S. M. (2003). *User Acceptance of Information Technology: Toward a Unified View USER ACCEPTANCE OF INFORMATION TECHNOLOGY: TOWARD A UNIFIED VIEW1* (tech. rep. No. 3).

Weber, S., Wyszynski, M., Godefroid, M., Plattfaut, R., & Niehaves, B. (2024). How do medical professionals make sense (or not) of AI? A social-media-based computational grounded theory study and an online survey. *Computational and Structural Biotechnology Journal*, *24*, 146–159. https://doi.org/10.1016/j.csbj.2024.02.009

Wen Loh, H., Ping Ooi, C., Seoni, S., Datta Barua, P., Molinari, F., & Rajendra Acharya, U. (2022). *Application of Explainable Artificial Intelligence for Healthcare: A Systematic Review of the Last Decade (2011-2022)* (tech. rep.).

Wright, K. B. (2006). Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, *10*(3), 00–00. https://doi.org/10.1111/j.1083-6101.2005.tb00259.x

Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2022). Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making. https://doi.org/10.1016/j.artint.2022.103839

Yousef Shaheen, M. (2021). Article title: Applications of Artificial Intelligence (AI) in healthcare: A review Applications of Artificial Intelligence (AI) in healthcare: A review. https://doi.org/10.14293/S2199-1006.1.SOR-.PPVRY8K.v1

Appendix A

Measurement scales

The questions to assess knowledge on AI used by Weber et al. (2024) consists of the following items.

Compared to the average person. . . .

1. I know much about AI-based technology.
2. I am very familiar with AI-based technology.
3. I am very interested in AI-based technology.
4. I use AI-based technology a lot.

All items were rated on a five-point Likert scale.

Appendix B

AI attitude scale

The questions to assess the attitude towards AI proposed by Grassini (2023) consists of the following questions. All questions are answered using a five-point Likert-scale.

1. I believe that AI will improve my work.
2. I think I will use AI technology in the future.
3. I think AI technology is positive for humanity.

Appendix C

Explanation Satisfaction Scale

The explanation satisfaction scale by Hoffman et al. (2018) consists of the following questions. All questions are answered using a Likert-scale with the following items: I agree strongly (1) - I agree somewhat (2) - I'm neutral about it (3) - I disagree somewhat (4) - I disagree strongly (5)

1. From the explanation, I understand how the model works.
2. This explanation of how the model works has sufficient detail.
3. This explanation of how the model works is useful for decision-making.
4. This explanation of the model shows me how accurate the model is.
5. This explanation lets me judge when I should trust and not trust the model.

Appendix D

Trust Scale for XAI

The trust scale for XAI by Hoffman et al. (2018) consists of the following questions. All questions are answered using a Likert-scale with the following items: I agree strongly (1) - I agree somewhat (2) - I'm neutral about it (3) - I disagree somewhat (4) - I disagree strongly (5)

1. I am confident in the model.
2. I feel safe that when I rely on the model, I will get the right answers.
3. I am wary of the model.
4. I like using the system for decision making.

Appendix E

Measurement scale Behavioral Intention

The questions to assess behavioral intention were adapted from Venkatesh et al. (2003). They were measured using a 5-point Likert-Scale.

1. I predict I would use the XAI system, if it is available to me.