UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

English
Logo

Director: Prof Dr. rer. nat. Michael Herczeg

Explainable Artificial Intelligence: Designing human-centric assessment system interfaces to increase explainability and trustworthiness of artificial intelligence in medical applications

Master's Thesis

as part of the study program
Media Informatics
of the University of Lübeck

submitted by:
Philipp Dominik Bzdok

Issued and supervised by:
Univ.-Prof. Dr. rer. nat. Thomas Franke, Dipl.-Psych.

with support from:
Tim Schrills, M.Sc.

Lübeck, 27th December, 2021

Acknowledgement

Kurzer Dank an Personen, die Sie bei der Arbeit unterstützt haben. Z. B. inoffizielle Betreuer, Teilnehmer an den Evaluationen (nie namentlich nennen), Medientechnik, Sekretärin, etc. pp. — nur wenn Sie den Personen wirklich dankbar sind. (Ist nett aber für die Bewertung irrelevant.) Falls nicht verwendet diese Seite einfach entfernen.

Text ...

Kurzfassung

Abstract schon für Zwischenabgabe schreiben. Später kommen dann noch Sätze dazu, aber Grundgerüst steht.

Kein "Teaser" sondern eine kurze Zusammenfassung (das, was man braucht, um sich schnell einen Überblick zu verschaffen, ob es sich lohnt, die Arbeit zu lesen).

Inhalt umfasst die zentralen Punkte aller Kapitel, von Ziel/Fragestellung bis Ausblick.

Nie länger als diese eine Seite (inkl. Schlüsselwörter).

Text ...

Schlüsselwörter

Verwendete Literatur gibt Hinweise auf passende Stichwörter. Das sind die Suchbegriffe, die man bei einer Literatursuche verwenden würde.

Text ...

Abstract

Englische Version der Kurzfassung. Nicht einfach Google Translate oder DeepL verwenden. Trifft die Nuancen nicht und klingt z. T. nach Yoda.

Text ...

Keywords

Text ...

Contents

1 Introduction

The use of modern Artificial Intelligence (**AI**) techniques is pervasive and can be found in many fields of application, such as digital image processing, search engines and speech recognition (European Commision, 2020). Other application fields, such as medical diagnosis systems, cannot benefit as easily from AI based technology compared to recreational domains. This impediment stems oftentimes from the AI being a "black box". In consequence humans struggle to understand such AI-systems and their output, leading to trust and compliance issues (Adadi & Berrada, 2018). These issues are further enhanced in the medical context where decisions, possibly based on AI, can have severe consequences for users, especially patients.

An example for medical human-AI-interaction is the image-based recognition of Deep Vein Thrombosis (**DVT**) with real time AI support for medical professionals by *ThinkSono*. The system leverages AI to guide the user through the current gold-standard diagnosis, a compression ultrasound examination, so that it enables any healthcare professional to detect DVT (ThinkSono, 2021). Closely related in this context is the interdisciplinary research project *CoCoAI*, which aims to explore psychological, ethical and technological implications of human-centered, AI based applications in the DVT diagnosis and beyond (CoCoAI, 2021).

When AI based systems are used in high risk application contexts, such as medical diagnosis, the aspects of explainability, interpretability and trustworthiness become a primary concern for adoption and use of said system. Ribeiro et al. (2016) already explored the importance of explainability and trust in AI based systems and postulated that AI systems will not be used if the users have no trust in the model or the results. Even though many machine learning algorithms score high on standard performance metrics, such as precision, recall or Area Under the Receiver Operating Characteristics (**AUROC**), user-facing performance may be way worse (Gordon et al., 2021). Understanding the AI's underlying machine learning (**ML**) model and its predictions is an important step for assessing trust and facilitating effective interaction (Ribeiro et al.,

2016). Recent technological advances are realized by *Clearbox AI*, with the focus on trustworthy AI by implementing an AI model assessment (Clearbox AI, 2021b; European Commision, 2021). The model assessment can help model owners to identify robustness issues, potential undesired behaviour, and explain errors and uncertainties regarding the model predictions (Clearbox AI, 2021a).

Trust in AI systems is primarily induced by the users' understanding and the general interpretability of the machine learning model and their predictions (Ras et al., 2018; Ribeiro et al., 2016). The wide array of different possible user groups and the complex constructs of understandability and trust demands for a human-centric approach in designing AI assessment systems. Because of the inherent complexity of non-linear machine learning models, especially Deep Neural Networks (**DNNs**) for image processing, suitable visualization and communication techniques are non-trivial. Additionally to the complex models for image classification, the input data is also more complex as it is unstructured. Non-linear neural networks and unstructured data provide additional challenges for Explainable Artificial Intelligence (**XAI**), as described in Keane and Kenny (2019). XAI is a research field that studies how AI decisions and data driving those decisions can be explained to people in order to provide transparency, enable assessment of accountability, demonstrate fairness, or facilitate understanding (Arrieta et al., 2019). XAI plays an important role in the acceptance and finally in the usage of AI based technology. This is further underlined in the medical context where public authorities set strict regulations on the usage of technological systems and ethic concerns have to be thoughtfully addressed.

In the context of a image-based medical diagnosis system, it is important that the responsible stakeholders, such as medical practitioners, specialized doctors and clinic managements, are enabled to make informed decisions on the usage of AI based technology, even though their expertise in machine learning and data science is expected to be low. The stakeholders' trust in this system is a primary factor for the widespread use of said technology for real life applications. Therefore, increasing the understanding of the AI model and finding an optimal trust level in the predictions by designing human-centric explanation techniques within the AI model assessment system is a main goal of this work. Additionally it is conceivable that authorities will instantiate auditors for AI based systems in medical contexts. Having a comprehensible and scientifically proven assessment system could be a big step in the approval and adoption of said system.

1.1 Goals

The users understanding of the AI model and trust in the model are highly essential as pointed out by Knapič et al. (2021). This holds especially true for medical applications where re-traceable results have to be provided and people acting on these results bear great responsibility. To facilitate understanding and trust the machine learning model has to be interpretable and explainable. In the context of Convolutional Neural Networks (**CNNs**) interpretability of models can pose a significant concern because of their inherent complexity. Explanations of AI models can provide insights on the machine's decision process and therefore generate user understanding. This can lead to the model being more interpretable by humans.

Assessing the suitability and performance of a CNN for a specific task by applying standard performance evaluation metrics is problematic, since these can be oblivious to distinguishing the diverse problem solving behaviors of a neural network (Lapuschkin et al., 2019). Lapuschkin et al. (2016), Ribeiro et al. (2018), and Samek et al. (2021) give an overview on the technical foundations of XAI and a presentation of practical methods, which will be used in conjunction with human-centric design to explore and evaluate suitable and efficient methods to explain a model's classification.

The goal of this thesis is to design, develop and evaluate interactive AI assessment system artifacts for medical professionals and machine learning specialists in a human-centric fashion to facilitate explainability and trustworthiness of AI models. Developing such a system, with human concerns in focus, leads to following research questions:

Q1: How is the stakeholders' (medical professionals, clinic managements or data scientists) subjective information processing awareness linked to trust for a specific model and its predictions in the medical domain?

Q2: How can different explanation techniques, ranging from perturbation based approaches to explorative alternatives, increase trust in image classifier models and predictions?

Q3: What are interaction methods to explain and possibly optimize trust levels in image classification models?

Q4: To what extend can structured metadata increase the stakeholder's understanding of a model's operational range and performance?

1.2 State of the Art

An AI assessment system is currently offered by Clearbox AI. The *AI Control Room* cloud platform enables users to assess, improve and validate ML models and data in accordance with the principles of Trustworthy AI (Clearbox AI, 2021b; European Commision, 2021). Clearbox AI (2021b) describes its AI assessment system as a "Deep Pre-production Analysis" tool:

> "AI Control Room automatically generates a model assessment to help model owners to identify robustness issues, potential undesired behaviour, and explain errors and uncertainties regarding the model predictions."

Concretely the product enables users to perform following tasks for AI models working on tabular data:

Model behaviour validation: Validation metrics and plausible causes of error are clearly presented, potential limitations and irreducible uncertainty are identified and local explanations of the model behaviour are generated selecting representative points in the data set.

Synthetic data generation: A generative model can be used to create synthetic data points that preserve the statistical properties of the original data set. These points can augment the original training set to improve generalization, to increase model robustness, and to oversample specific labels when in the presence of unbalanced data.

Data-centric analysis: Generative models perform a probabilistic analysis of the underlying data allowing for robust outliers detection and uncertainty analysis. This information can help you to evaluate data quality.

Centralised tracking system: AI Control Room acts as a centralised tracking system to store lineage, versioning, and metadata of your data sets and models. Assessments generated are securely persisted along with models and data sets.

Besides general information and standard metrics (see Figure 1.1) the assessment system offers varied insights into different aspects of the machine learning model: Figure 1.2 shows graphs of training and validation precision, recall and calibration, while Figure 1.3 shows the models strong points and limitations by analyzing the feature distribution in

the data. Furthermore, the second half of the model assessment focuses more on the interpretability aspect of machine learning: Figure 1.4 shows a confusion matrix of possible classification results, which is then extended by example data points, chosen by the assessment system (see Figure 1.5). These examples can then be further explored to generate understanding of the models inner workings by applying an attribution based explanation technique combined with a decision rule explanation.

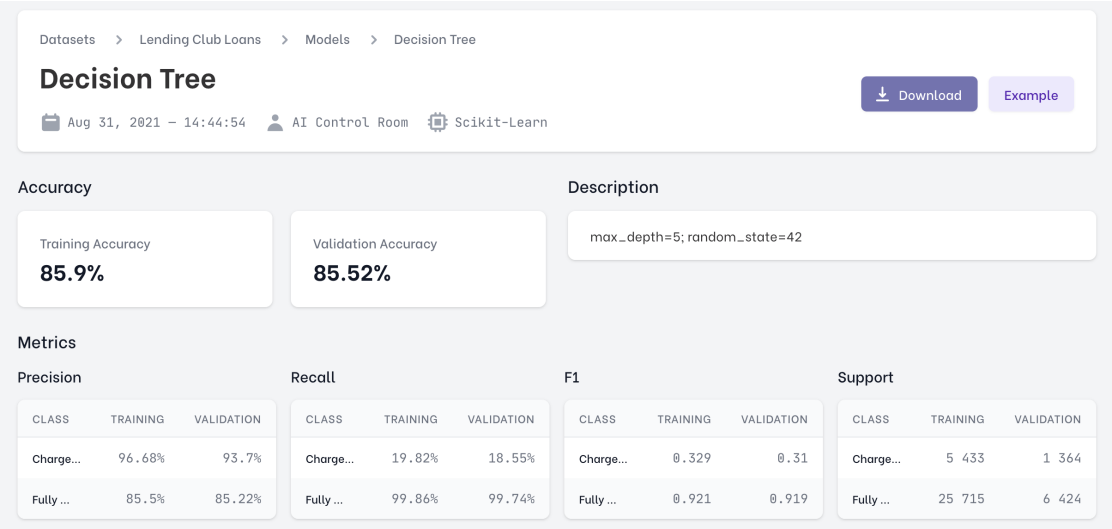


Figure 1.1: AI Control Room - Model Assessment Overview with Standard Metrics

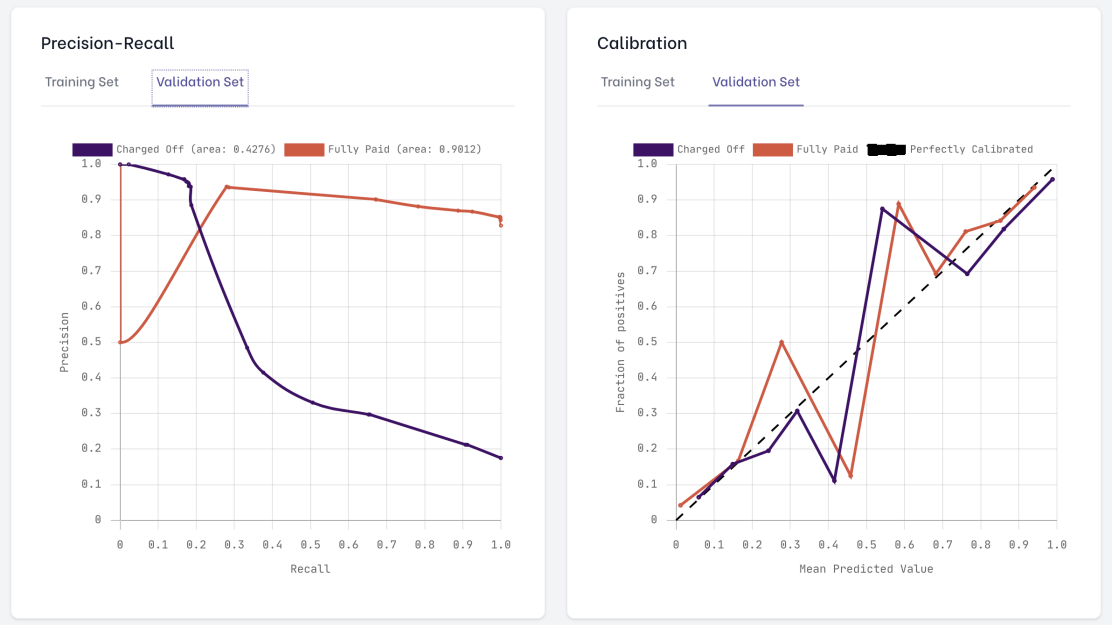


Figure 1.2: AI Control Room - Precision-Recall and Calibration Graphs

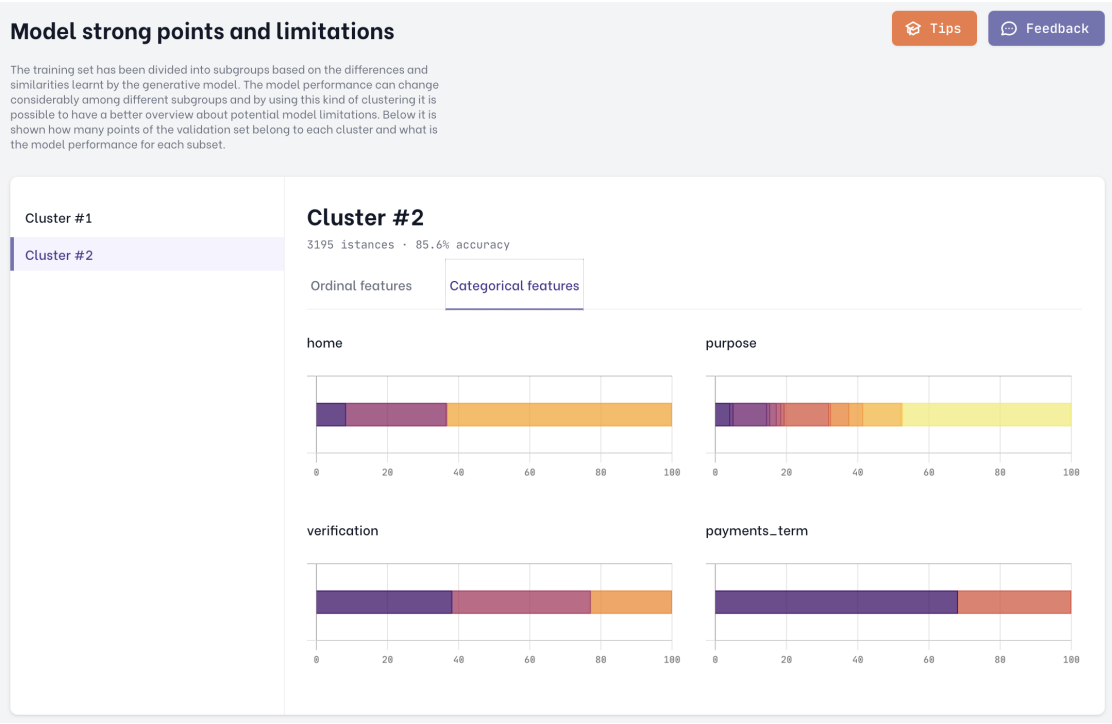


Figure 1.3: AI Control Room - Model String Points and Limitations

Figure 1.4: AI Control Room - Interpretability Assessment

HOME	GRADE	PURPOSE	INQUIRIES	DELINQUENCY	LOAN_AMOUNT	BANKRUPTCIES	FICO_AVERAGE	VERIFICATION
Mortgage	2	debt_consolidation	4	0	10000	0	732	Not Verified
Mortgage	4	other	5	5	8000	0	667	Source Verifi
Rent	6	other	5	0	18000	0	677	Verified
Mortgage	5	credit_card	8	0	12500	0	677	Verified
Rent	7	debt_consolidation	3	0	5875	0	642	Not Verified
Mortgage	3	credit_card	5	0	19000	0	722	Verified

Figure 1.5: AI Control Room - Example Data

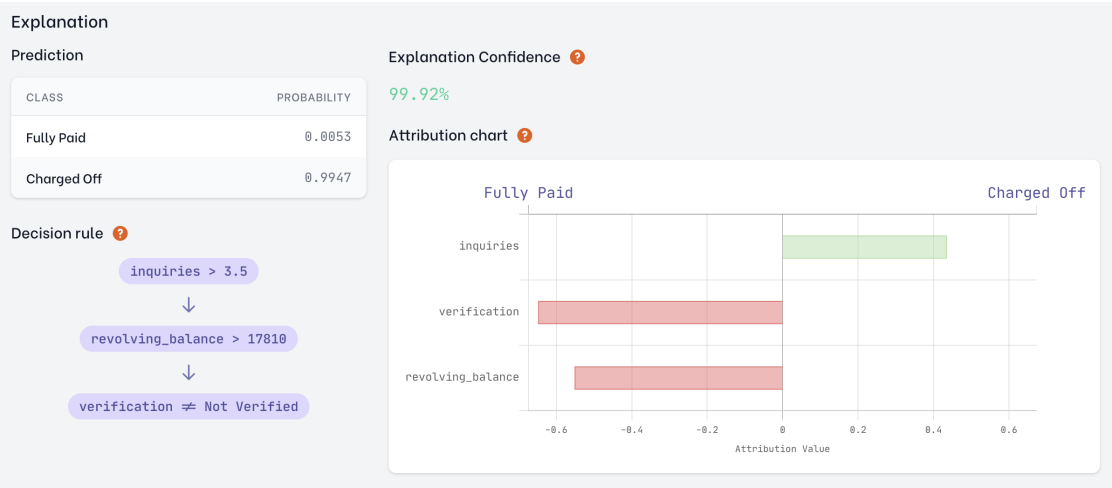


Figure 1.6: AI Control Room - Prediction Explanation for Examples Data

1.3 Approach

As already described in the introduction of chapter 1, many machine learning algorithms score high on standard performance metrics, but user-facing performance may be way worse (Gordon et al., 2021). This issue is caused by real world applications being very dependent on the actual human-AI-interaction. Following this reasoning, it lends itself to utilize a human-centered design process for creating AI assessment systems. Figure 1.7 shows a standardized process of human-centered design, which was applied in this thesis to conceptualize, implement and evaluate assessment system artifacts. The key take-away is the inclusion of human aspects in all stages of the process. Research on evaluation of AI explanations revealed that there is a big gap between the perceived and actual usefulness of explanations, as described by Ras et al. (2021). This further underlines the need for a human-centered approach in designing AI assessment systems.

The thesis' structure will reflect the human-centered approach, which is visualized in Figure 1.7: As already alluded in chapter 1, chapter 2 is about understanding and setting the context of use by conducting literature research and user interviews. Based on the established requirements chapter 3 will describe the conception of functionalities and interaction design. The development of solutions will be described in chapter 4, while chapter 6 is about the evaluation of the solutions. This general process is embedded in a iterative loop, where intermediate results are evaluated against the requirements and subject to change.

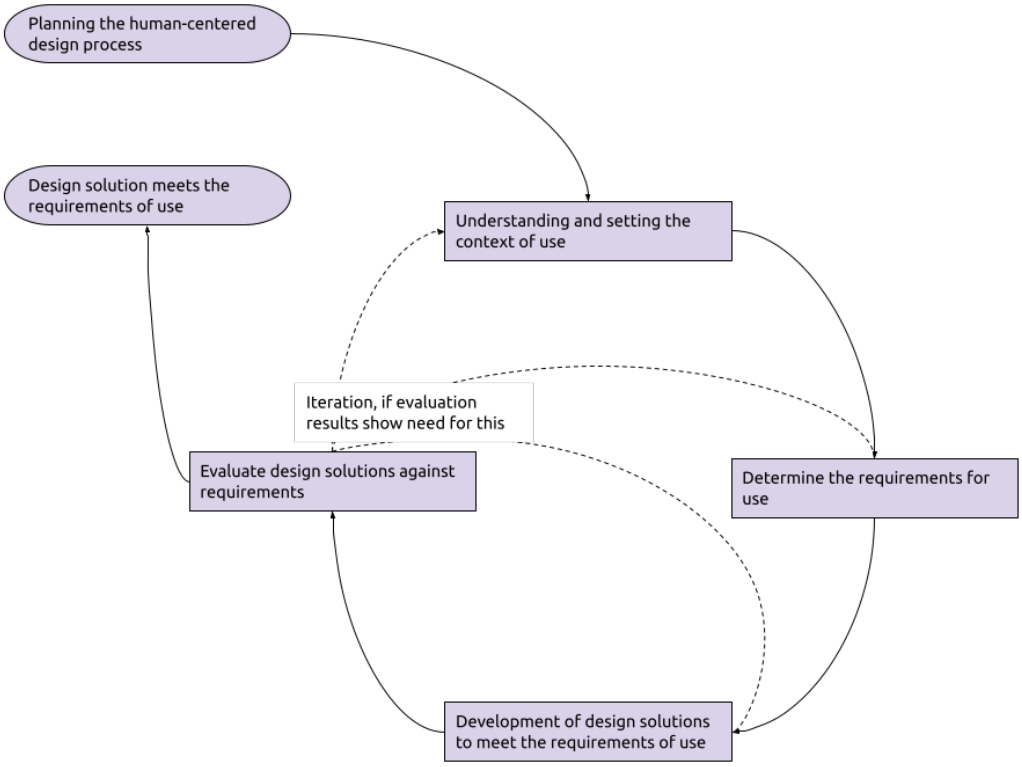


Figure 1.7: Human-centered Design Process (*DIN EN ISO 9241-210*, 2011)

2 Analysis

Following the human-centered design process, it is important to incorporate the potential users from the beginning (Harte et al., 2017). This is also reflected in the analysis, where the context and setting of use has to be understood and set. Besides the human factors, there are also more general and theoretical aspects to be analysed, such as the context of AI in the medical application domain for specific tasks, such as classifying disease patterns with medical imaging.

In the following the context, task, problem and users will be described and analysed as a foundation for the human-centered design process and the following conception of AI assessment system artifacts.

2.1 Data Sources

Three main data sources where used for the analysis, ranging from general scientific literature about XAI to expert interviews and cooperation with developers of an existing application.

2.1.1 Scientific Literature

As described by Mueller et al. (2019), the amount of scientific publications on the topic of explanation in intelligent systems has surged in the last 5 years, revealing many important and relevant information on this subject area through openly accessible papers. In the beginning of the thesis (July 2021) a general search on *Google Scholar* was conducted to gain a overview on available publications. A non-exhaustive list of search terms at the time was:

- XAI

- XAI in Medical Applications
- Explainable Artificial Intelligence
- Explainable Artificial Intelligence in Medical Applications
- Explainable Machine Learning
- Interpretable Machine Learning
- Explaining Black-Box Machine Learning Models
- Explaining DNNs

This general research yielded already good results, as there were many relatively new and popular publications on the topic of XAI, such as Adadi and Berrada (2018), Hoffman et al. (2019), Mueller et al. (2019), and Ras et al. (2018), Ras et al. (2021), dealing with (1) issues of black box models (2) trustworthiness and transparency of AI-based systems (3) challenges of AI explanations (4) synopsis of key ideas of XAI (5) compilation of existing XAI methods among other topics.

The results of the internet research were then further reinforced by academic partners from the University of Lübeck, with whom related research was conducted in the context of the *CoCoAI* project. Additional important scientific papers for the analysis were: Arrieta et al. (2019), Knapič et al. (2021), and Samek et al. (2021), dealing with (1) review of AI explanation methods (2) XAI methods for decision-support in medical image analysis scenarios (3) literature analysis of XAI publications.

The literature research resulted in a foundation on the XAI topic with a thematic focus on black box image classifiers, available explanation methods and trustworthiness of AI applications. While many of the presented publications discuss theoretical and technical aspects of XAI, human-centered perspectives were found to be rare. Samek et al. (2021) for example gives a detailed theoretical and technical overview of many explanation methods, but reduces the human interpretability of visual explanations to only the file size, concluding that lower file sizes are more interpretable to the user. Furthermore Ras et al. (2021) presents many ways to objectively compare explanation methods, but also highlights the difficulties of user-centered explanations, tailored to specific requirements such as medical professionals have for decision support systems. However there is also literature on psychometric evaluation of XAI systems (Hoffman et al., 2019) with a strong

focus on human aspects. Overall the literature research showed that the XAI domain heavily focuses on the theoretical, technical and organizational aspects of explanation methods, while only very few authors consider user-centered research methods from the HCI domain, which also found by Mueller et al. (2019).

2.1.2 Interviews

Complementing the general research on XAI, expert interviews where conducted. These interviews specifically target medical professionals and data scientists. Interviews with the actual user group of a potential solution is key to understanding and setting the context and requirements of use. The participants for the interviews were chosen with following requirements in mind:

Medical Professionals: Has interest and/or knowledge in AI-systems; has worked with or researched AI-systems in the medical domain; can judge the benefits and risks of the use of AI in medical applications.

Data Scientists / AI Researchers: Is familiar with the XAI topic; has interest in explainability and trustworthiness of machine learning models; has worked with AI in the medical context.

Participants for the interviews were gathered via academic partners, internet research and word of mouth. 16 suitable people from 6 different institutions were contacted about potential interviews. Ten leads were medical professionals while six leads where data scientists or AI researchers. In total, 4 interviews were conducted. The participants are described in further detail in Table 2.1.

The Interviews were conducted in German and executed as one-to-one online interviews. The interviews were recorded for later reference if consent was present. Additionally the interviews were supported by a research colleague, who kept protocol. After the interviews the recordings were transcribed for further analysis and the participants were asked to answer the *Affinity for AI Interaction* (**AAII**) questionnaire, which is a modified version of the *Affinity for Technology Interaction* (**ATI**) questionnaire (Franke et al., 2019). ATI aims to determine the tendency to actively engage in intensive technology interaction, as a key personal resource for coping with technology. Analogously the AAII questionnaire aims to determine the tendency to actively engage in AI interaction. The

ATI questionnaire was modified to focus on AI interaction for a better thematic fit, as it is important to estimate how the user likes to deal with AI.

In terms of content, the interviews for medical professionals and data scientists were slightly different as seen in Table 2.2. The reason for this is the heterogenous expertise on the subject area of machine learning models and potential user requirements. The whole interview guideline can be found in Appendix B: Interview Guideline.

ID	Age	Gender	Occupation	Education Level	AAII Score
1	28	male	Assistant Physician in Neuroradiology	State Examination	4.89
2	24	female	Research Associate (ML)	Masters Degree	5.12
3	48	male	Surgeon	Dr. med.	5.45
4	27	female	Assistant Physician in Neuroradiology	State Examination	4.67

Table 2.1: Interview Participants

Medical Professionals	Data Scientists / AI researchers
Actual Usage of AI	Actual Usage of AI
Perspective on AI Usage	Comparison of AI Models
Trust in AI	Perspective on AI Usage
Potential Problems with AI Usage	Trust in AI
Own Explanation Techniques	Potential Problems with AI Usage
Familiarity with XAI	Own Explanation Techniques
Assessment on Local Explanations	Familiarity with XAI
Information Processing	Need for Local Explanations
Reliability vs. Trust vs. Understanding	Need for Global Explanations
	Information Processing
	Trust-Behavior Connection
	Reliability vs. Trust vs. Understanding

Table 2.2: Interview Topics

After gathering the interview data via protocols and transcripts a thematic analysis according to Braun and Clarke (2006) was applied to identify common topics and codes. The thematic analysis is a widely used qualitative analysis method mainly found in the field of psychology and can be used as a primary tool to access data from interviews. Applying this method resulted in a thematic map showcasing common overlapping topics found in the interviews, which can be seen in Figure 2.1. The thematic map serves as the baseline for further context and requirement analysis.

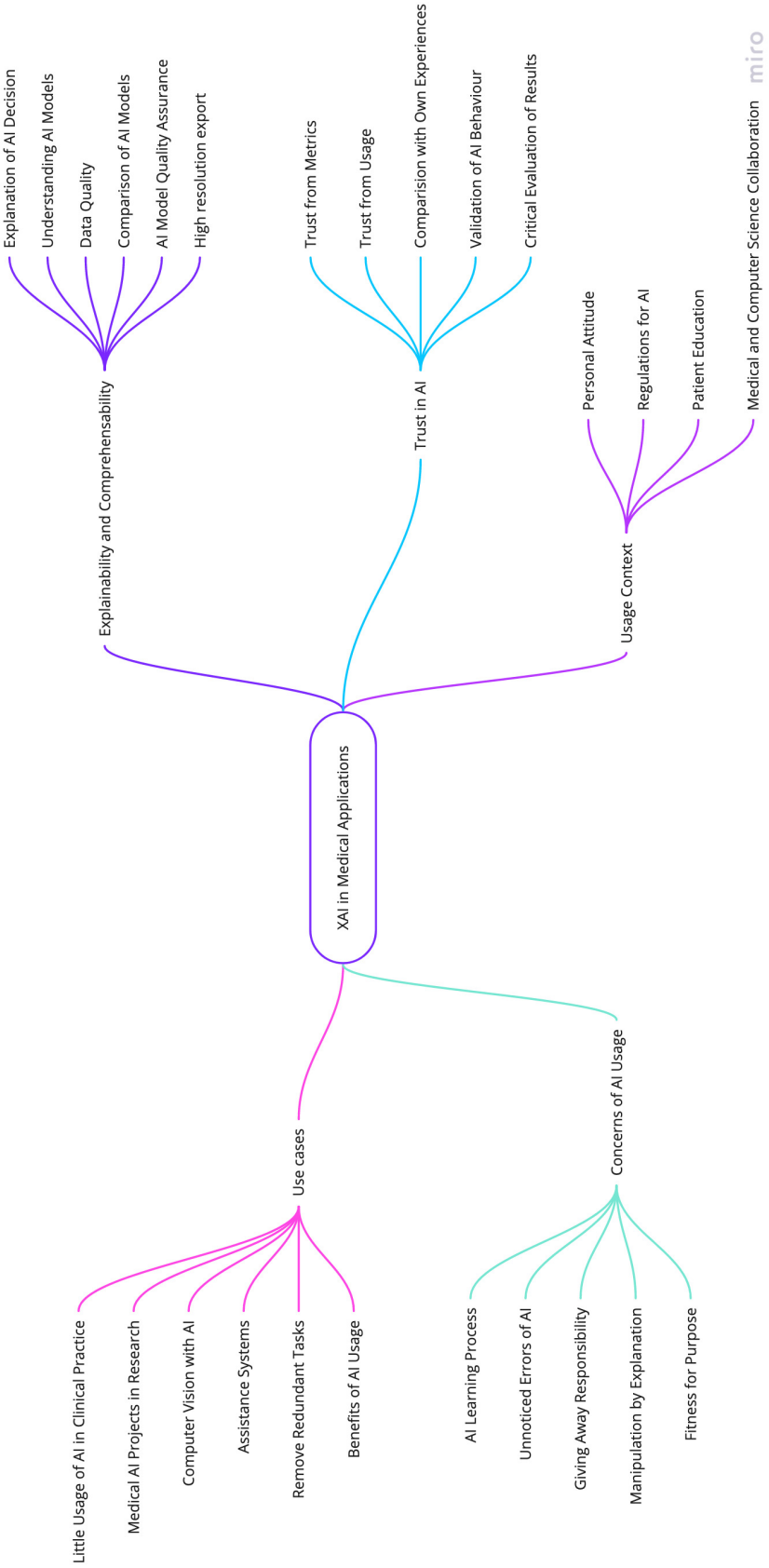


Figure 2.1: Thematic Mind Map

2.1.3 Existing Applications

A scientific cooperation with Clearbox AI allowed us to access a additional source of information, valuable for the analysis and conception. As already described in section 1.2 Clearbox developed an AI model assessment solution among other things. During the period of cooperation regular meetings were held with the CTO and other employees. These meetings were used for knowledge exchange on the subject of XAI literature, previous experiences, feedback and conceptional workshops. In particular, resources such as Facebook (2021), Google (2021), and Streamlit Inc. (2021a) were valuable additions to the analysis. Furthermore the (beta) access to the Clearbox AI *Control Room* cloud platform and the communication of user feedback was also used to gather information for the analysis and conception of an assessment system for image-based AI models.

2.2 Context Analysis

Black-box DNNs have become pervasive in todays society and represent a proven and indispensible machine learning tool. While these machine learning models can easily be used in recreational non-risky contexts, this does not hold true for the medical domain where decisions based on the results of a machine learning model can bear great risks for users and patients. This issue stems from the lack of interpretability and trustworthiness of DNNs (Adadi & Berrada, 2018). DNNs architectures are inherently hard to understand and therefore interpretability of and trust in the results of such neural networks are a challenge.

Creating a solution to explain AI models a priori to the use can help setting the right expectations towards the AI model. Consequently the users of such a explanatory system gain the ability to build a fair mental model before using the AI, which in turn can support the formation of appropriate levels of trust (Hoffman et al., 2019). This is beneficial to the user and facilitates efficient usage of the model in production (Google, 2021; Hoffman et al., 2019).

Explaining the model a priori also enables the solution to build computationally complex explanations, which can depend on data sets of thousands of images. Supplying the user with explanations potentially based on the whole data set can also be beneficial: Statistical analysis and clustering of the data and metadata can support the under-

standing of model limitations and edge cases, while exemplary local explanations can be statistically distributed to gain a better overview of global behavior of the model. Combining different types of a priori explanations improve the coverage of the key attributes of explanations - explainability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness - which were described by Hoffman et al. (2019).

The aspects mentioned above are also reflected in the interview data. The following list showcases with *DeepL* (DeepL GmbH, 2021) translated quotes from interview partners regarding the need for (a priori) explanations for AI models aimed at the medical domain:

> "I think you have to be critical and look at the results carefully - is the result at all plausible?"
>
> "What is the information that is interesting for the system or that is decisive for the decisions? What information is rather irrelevant?"
>
> "If the explanations for the model are based on any data that are meaningless from my clinical experience, such as a stroke being pinned down by bone shapes. That would shake my confidence in the machine, even though it might give reliable results."
>
> "If, for example, there is an outlier data point in a specific case and you are not quite sure why it is like this or how you should interpret it, then something like this [local explanation] is great, so that you can understand why the result is like this or like this."
>
> "I've always been interested in exactly how this works, how much training data it's based on, what's behind it, why the system decides the way it does."
>
> "[...] and you might also learn what the machine pays attention to, which I personally could also pay attention to when I look at the picture. That would certainly strengthen my confidence in the application."

Overall the interview partners expressed a strong need for explanation and validation of AI models in medical applications. This finding highlights the need for explanations for the users of medical AI, where they are able to explore the data and capabilities of the AI model and therefore build appropriate levels of understanding and trust. Fur-

thermore the medical professionals homogeneously stated a need for validation of the AI model's results with their own expertise prior to the clinical usage of the AI. Therefore an interactive a priori explanation tool implemented in a sandbox environment could satisfy the needs expressed by the professional users for their usage context.

2.3 Problem and Task Analysis

The following problem scenario summarizes the starting point of the problem and task analysis: The general medical practitioner Dr. med. Mustermann wants to offer thrombosis diagnosis in his office. He is not specialized in vein examination and thus has just basic knowledge and also no necessary equipment. In the past this has lead to him referring patients with venous disorders to a specialized clinic. Through a colleague he was made aware of "*AutoDVT*", a AI based software developed by ThinkSono, which can help him offer DVT diagnosis in his office (ThinkSono, 2021). The AutoDVT system works with image-based machine learning and can support medical professionals in real time with identifying DVT. Since Dr. med. Mustermann has very little knowledge of AI and machine learning he is very skeptical towards this innovative but foreign software system. Although he sees the immediate benefits of using a system which supports him with the examination of DVT, his trust in the system's predictions is very low and he fears relying on the AI's assessment. AI based technology is a black box for him, which he does not fully understand. The predictions of the systems are opaque to him and thus lead to rejection of the system.

DNNs for image classification are able to detect various disease patterns with medical imaging and can be used by medical professionals to support diagnosis and possibly increase efficiency and effectivity if trusted and used correctly (Adadi & Berrada, 2018; Knapič et al., 2021). However, the reality looks different: Medical professionals bear the responsibility for their decisions regarding the patient and thus are rather relucant about using AI based systems - even though the AI could outperform them in image classification tasks. Most of the time decisions are based on personal experience, which was developed over a long period of time. This sentiment is reflected in the interview statements of medical professionals:

> "One risk, I believe, is also that you hand over responsibility to the machine."

> "The classic risk of simply relying on what the algorithm says."

> "The doctor with his expert knowledge will always compare this with his knowledge and experience, is this correct, what is the probability, is this in the range?"

Modern DNN based image classificators, such as the `XRV-DenseNet121-densenet121--res224-all` model from Cohen et al. (2020), can provide very good results in the prediction of pathologies. For example, the prediction accuracy for pneumonia is benchmarked as 86% (Cohen, Viviano, et al., 2021). The use of such model or comparitive ones could benefit medical professionals in many ways as described by the interview partners:

> "It [AI system] facilitates standardized findings in particular"

> "It [AI system] might give you a little peace of mind that you haven't missed anything."

> "I always think to myself that this is based on CT gray levels, i.e. on these density values, and I always think to myself that it makes sense that a computer can distinguish these density levels better than my eyes."

The benefits of using a DNN based image classificator to support medical professionals in detection and diagnosis of diseases are clear. For users to leverage these benefits trust in such a system must be given, which cannot be generated by mere accuracy metrics (Samek et al., 2021). To overcome the hesitation of using AI in medical applications, a AI assessment system can be used to facilitate understanding of and trust in the algorithms. Stakeholders, such as practitioners or clinic managements, can use such a system to gain customized insights into the model and data prior to using it in everyday clinical practice.

2.4 User Analysis

AI explainability and trustworthiness must be considered with respect to the individual who is regarded as the beneficiary of the explanation (Mueller et al., 2019). For an assessment system that aims to explain AI models in the medical domain those user groups are: (1) Medical professionals, such as practitioners and clinic managements and (2) data scientists and AI researchers developing AI models. Naturally those two groups

differ greatly from each other. While medical professionals have great expertise in various fields of examination, diagnosis and communication of pathologies, they also are expected to have little knowledge in computer sciences and machine learning. Data scientists on the other hand do have great knowledge of computer sciences, machine learning and neural network architectures, but lack the concrete medical expertise. Following the human-centered process the needs, requirements and whishes in regard to an AI assessment system of those user groups were analysed, mainly referencing the interviews from subsection 2.1.2 and the resulting thematic map (see Figure 2.1).

2.4.1 Medical professionals

Through the interviews it was found that the surveyed participants differed in their experience and disposition towards AI systems. Surveying practitioners of different ages showed, that especially the younger ones, working in neuroradiology, are open towards using AI in their daily routine or that they are already using it. Some participants, took special courses on machine learning in the medical domain during their studies. Then again older, more experienced participants stated much less contact with AI in clinical practice. All interviewed medical professionals showed interest in medical AI in research projects and were positive on the benefits of AI, especially computer vision tasks. The most cited benefits were (1) great ability of machine learning to identify pathologies in medical images (2) take over of redundant tasks (3) backup for diagnosis and the handling of computationally expensive tasks. The medical professionals were also wary and timid about using AI. This disposition was stated to be mainly rooted in the missing experience in using and trusting such AI systems. The black box character of DNNs was stated to be a central issue: Not being able to re-trace the decisions of the AI and having to give away responsibility lead to trust and compliance issues, which was already stated by Ras et al. (2021). Depending on standard metrics was also stated to be insufficient and the interviewed experts showed interest in the training data set and active comparison of the AI's results with their own experiences. While standard metrics, such as accuracy, sensitivity and specificity were important to the interviewees, the critical evaluation of the results and the validation of the AI behavior were whished for by every one of them.

As Table 2.1 shows, all interviewees had a high affinity for AI interaction. This is a important aspect to consider, since it shows their tendency to actively engage in interaction with AI while also being interested in it. This fact explains that the physicians were so

interested in the explainability and comprehensibility of AI models. The interviewees stated that the explanation of AI decision and therefore the understanding of the model is important to them. Also the training data and its quality was a very common topic amongst all participants. Interestingly it was also stated, that understanding and trusting the model is important to being able to propagate the knowledge and trust to fellow medical practitioners and also patients.

Even though the interest in the functionality of machine learning models was big, the medical practitioners admitted that they have little knowledge on this subject and are limited regarding understanding the technical complexities. However they also stated that there is ongoing collaboration with AI researchers and software engineers for research purposes.

2.4.2 Data scientists and AI researchers

This potential group of AI assessment system users stand in great contrast to the previously mentioned one. Data scientists and AI researchers have a good understanding of the complexities and inner workings of machine learning algorithms. Therefore the requirements and needs of this user group are expected to be very different from the medical professionals. As Table 2.1 shows, unfortunately only one AI researcher could be interviewed during the analysis.

The interviewee stated experience with many kinds of neural networks, while also being familiar with clustering, featuring and interpretation tools. The perspective of AI researchers on interpretability and trustworthiness seems to be also quite different. Important aspects mentioned were: Performance metrics do not guarantee usability in real applications; separation of system results and system architecture; experimental validation; relations to developers; reviewing code. Understanding the complexities of such AI systems was a key aspect as stated by the interviewee. For this literature and study experience were mentioned to be crucial. Even some experimentation with heatmap-based explanation tools were used to understand AI models better.

While data scientists and AI researcher are different to medical professionals in their expertise, some overlap was found in the interviews: The interviewee stated to generate trust by doing exemplary input-output experiments, screening the training data set and reviewing own expectations. Another common topic was the insufficiency of standard

metrics for assessing algorithm performance.

An explicit topic was the use of local explanations. The interviewee stated interest in local explanations as they are needed for improving their own understanding and for publications as proof that an algorithm works. Global explanations were not distinguished from local explanations by the interview partner, since the goal seemed to be the same: Determining if a model has weaknesses and to what it can be safely applied. Generating a benchmark for comparability of models to enable better adoption was wished for. Furthermore it was stated that theory should be researched further for understanding in addition to empiricism, which is also stated by Google (2021):

> "When is an explanation really meaningful? Explaining everything is difficult, but finding out when explanation should be given."

2.5 Conclusion on the Analysis

While the use of DNN based image classifiers for medical applications has many benefits, actual adoption is impeded by the black box character of such systems. The potential users, such as medical practitioners have trust and compliance issues. Even though the users see the immediate benefits of such AI based systems, especially in computer vision tasks, the issue with handing over responsibility to a system they do not understand prevails.

Insufficient standard metrics shall be supplemented with more interactive explanations with a focus on the comprehensibility of complex AI models. Promoting the formation of an appropriate mental model and therefore trust in the abilities of such systems is a key aspect which was identified by the analysis. The use of a priori explanations via an assessment system where the users can experiment AI models in a protected environment covers many requirements of the users: (1) Screening the training data (2) exploring model strengths and limitations (3) analyzing visual explanations for images and (4) doing input-output experiments. These are the core requirements for an assessment system to increase explainability and trustworthiness of AI in medical, image-based applications. Other relevant findings suggest that such an assessment system needs to actively consider the intention of the user, since it can vary greatly depending on the person's background: Medical professionals with low expertise in machine learning may need to have a more guided user experience, while experts on the subject of machine

learning may prefer a more open interaction style. Furthermore it is conceivable that statistical clustering, based on (training) metadata distributions can improve the explanation satisfaction by offering a balanced access to huge data sets in a way that is not susceptible to biases.

3 Conception

The conception of the AI assessment system builds on the requirements specified in the analysis. Based on the thesis objectives and user needs, functionalities have been derived. Core aspects to be adopted from the analysis are (1) interactive exploration of data (2) visual explanation of attribution (3) visual explanations (4) input-output experiments and (5) interaction guidance. Before diving into the details of functional conception (section 3.3), interaction design (section 3.5), and system architecture (section 3.4), an overview of the conceptual approach (section 3.1) and foundational use cases (section 3.2) will be given.

3.1 Conceptual Approach

Based on chapter 2 and the primary objectives, the conception follows the process described in Figure 1.7. To begin with, building upon the previously gathered information, a functional specification was created. This specification relies heavily on the thematic analysis of the interviews (see Figure 2.1). The first step of creating the functional specification was to identify user requirements and common use cases (J. J. Garrett, 2021; Kopetz, 1976) for an AI assessment system which can be used by medical professionals. The functional specification defines formal tasks, sub-tasks and required capabilities for those use cases. The formalized functionalities were then used as the foundation for an interdependency analysis, which highlights coactive human-computer-interaction (**HCI**) design patterns (Johnson et al., 2014). Having the functionalities defined allows for conception of interaction and interface design, accompanied by the technical system architecture. The interaction design was developed through flowcharts and interaction dialogues, building on the insights from chapter 2. The interface design concepts were created as non functional mockups that heavily referenced Google (2021) and Clearbox AI (2021b) design language. Having *Control Room* by Clearbox as a reference allowed for the leveraging of their expertise in conceptualizing an AI assessment system. A con-

ceptual iteration was performed by conducting an expert workshop on interaction design with research partners from Clearbox and the University of Lübeck. The results from interdependency analysis, interaction dialogues and interaction flowcharts were then realized in chapter 4.

3.2 Use Cases

Two common use cases found throughout the user requirements are to be presented as an anchor for further conception and reference. The main difference in these use cases is the underlying user group and therefore the intention of interaction.

3.2.1 Understanding through Interaction

A main use case emerges from chapter 2: Medical professionals who want to understand and trust machine learning algorithms through extensive interaction with the training data, visual explanations and comparisons before actually using the system in their daily clinical life. Meske and Bunde (2020) also highlights the importance of XAI methods for medical professionals in critical applications such as medical diagnosis systems. Medical practitioners or clinic managements are enabled to form appropriate expectations and understanding of the model's capabilities and performance by using XAI methods, which helps them adopting the model and propagating knowledge and trust to professional peers and patients.

3.2.2 Comparison of Models

An additional use case can be conceived, which revolves around users with a better understanding of machine learning: The interactive comparison of different AI models to be used in the context of medical application development. Using an AI model assessment beyond standard metrics can facilitate better decisions in favor or against a specific machine learning model. The ease of use and high interactivity of such system enables specialized users to explore more intricate facets of DNNs. Combining standard metrics with data exploration, visual perturbation techniques and the ability to experiment freely with the model creates a sandbox environment for testing and comparing

different machine learning models and data sets. Meske and Bunde (2020) also identified XAI methods to be beneficial for a better understanding of an AI model and therefore facilitate their validation and regulatory compliance.

3.3 Functionalities

Referencing section 3.2, interactive explanations which offer access to information about the training data and the model's functionalities, strengths and limitations will be the focus of the AI assessment system prototype. Furthermore visual and comparative explanations shall be supplemented to the presentation of image data, as described by Zeiler and Fergus (2013) and Cai et al. (2019). Considering the low expertise of medical professionals in machine learning topics, general explanations of system capabilities in textual form are expected to be beneficial. Additionally the ability to execute input-output experiments with data and the actual model were universally requested. Table 3.1 summarizes the specified functionalities.

The literature research from subsection 2.1.1 on different visualization techniques yielded many possible algorithms to further explain the ML model's reasoning. The main resources for this were Adadi and Berrada (2018), Arrieta et al. (2019), Ras et al. (2021), Ribeiro et al. (2018), and Samek et al. (2021). A wide selection of implementations for visual explanation techniques are available, ranging from perturbation-based to model intrinsic methods as described by Ras et al. (2021). From this lot of techniques, three were found to be particularly interesting based on their features. Table 3.2 shows the selections and the main properties of those. Moreover subsection 3.3.1 further characterizes all explanation methods that are thought to be suitable for the assessment system application.

The functionalities, more specifically tasks, as defined in Table 3.1 were subject to an interdependency analysis. The goal of the analysis was to define sub-tasks and capabilities required for those, which give insights into the mode of interaction between computer and user. Table 3.3 showcases the key aspects of the analysis in form of a interdependency analysis table as conceived by Johnson et al. (2014). The table makes it clear, that the human interacting with the system is highly dependent on the computer - this is no surprise in the context of a explanatory system, where the user seeks out information about a complex model. On the other hand the computer does not depend much on the

human counterpart, because it has the prevalence of information, but the computer can still benefit from the human expertise in certain situations. Overall this highlights the importance of interaction design and system collegiality for an AI assessment system.

#	Functionality	Description
1	Browse training data for given class	The user is able to explore the whole data set, which was used for training of the ML model. Additionally it is possible for the user to filter the data based on classification labels.
2	Show examples of false negative, false positive and low confidence	The user is able to explore training data for which the classification resulted in false negative, false positive or low classification confidence
3	Group data based on similarities	The user is presented with clustered training data that was algorithmically identified to be similar.
4	Show data that is very similar to data from another class	The user is presented with comparative explanations which showcase data points that are very similar to other data points but classified differently
5	Offer overview of general system capabilities	The user is presented with general information and metrics about the ML model in a written and structured way
6	Show written Explanations	The user is presented with written explanations about the system by leveraging text templates
7	Input-Output-Experiment	The user is able to feed the ML model with data and predict the models result. The model then also predicts a result, which is then compared with the user's prediction

Table 3.1: Functional Specification

3.3.1 Explanation Techniques

Three groups of explanation techniques are created as possible ways to implement the required functionalities (Table 3.1). The groups are divided by their style of interaction and information deliverance. Most of the presented techniques have a direct mapping to the functionalities to be provided by the assessment system, especially *General De-*

Method	Type	Description
Occlusion	Perturbation-based	Replaces rectangular areas in input with baseline reference and computes difference in output; Most useful in models where pixels in contiguous regions are likely to be correlated
Anchors	Rule-based	Shows part of the input (super pixels) which are sufficient for the classifier to make the prediction; Builds on the shortcomings of LIME (Ribeiro et al., 2016)
Layerwise Relevance Propagation	Model intrinsic	Propagates the prediction backwards using purposely designed local propagation rules; allows for differentiation between positive and negative influence of input pixels to prediction

Table 3.2: Visual Explanation Methods

scription, Metrics, Data Browsing and Experiment. The visual explanation techniques though, do overlap significantly in their practicality. While they differ drastically in their mode of operation and technical background, they do cater to the same interaction: Showing image regions that are significant to the result of the AI. Because of this and the limited time and resource frame, only one of the three proposed visual explanation types will be implemented.

Textual

General Description: General descriptions of the system capabilities aim to explain the system on a level, that is accessible to machine learning laymen, such as medical professionals. As described by Google (2021) it is beneficial for building trust when the system capabilities are explained instead of the technology itself. This helps the user to build a better mental model, especially when dealing with hyper-complex structures, such as DNNs. The general descriptions shall contain a textual explanations of the model capabilities to ensure a good introduction and appropriate expectations.

Metrics: Standard metrics are also used to describe the machine learning model. The metrics used are *Accuracy, Precision, Recall & F1.* Those metrics were chosen, because they are often used to describe other processes and applications in the medical context and therefore should be familiar and interpretable by the user.

Table 3.3: Interdependency Analysis Table

Tasks	Sub-Tasks	Required Capabilites	Team Member Role Alternatives				Interpretation and ODP Requirements
			Alternative 1		Alternative 2		
			Performer	Support	Performer	Support	
			Computer	Human	Human	Computer	
	Choosing class	Knowing classes					Either can choose the class. Assistance by the computer could improve the reliability of the human. The human may not know all classes thus requires directability.
	Access to data	Having storage					Only the computer can perform this task
Browse train data for given class	Display of data	Modify interface					Only the computer can perform this task
	Find data with given characteristic	Filter data					The computer could improve reliability by being supported by the human. Thus observability and directability is required. The human can only tediously perform this task and would require assistance.
Show examples of false pos. / neg. or low confidence	Display data	Modify interface					Only the computer can perform this task
	Choose characteristic	Know characteristics					Either can perform this task but would benefit in reliability from assistance. Thus directability is required.
	Find data with given characteristic	Filter data					The computer could improve reliability by being supported by the human. Thus observability and directability is required. The human can only tediously perform this task and would require assistance.
Group data based on similarity	Display data	Modify interface					Only the computer can perform this task
	Choose class	Know classes					Either can choose the class. Assistance by the computer could improve the reliability of the human. The human may not know all classes thus requires directability.
	Find data with high similarity but different classification	Filter data					The computer could improve reliability by being supported by the human. Thus observability and directability is required. The human can only tediously perform this task and would require assistance.
	Find data with correct classification	Filter data					The computer could improve reliability by being supported by the human. Thus observability and directability is required. The human can only tediously perform this task and would require assistance.
Show data that is classified as one class but very similar to data from another class	Display data	Modify interface					Only the computer can perform this task
Offer overview of general system capabilites	Translate system specifications for laymen	Having overview of specifications					The computer knows the systems specification but may need assistance to translate it, thus requiring observability and directability. The human can only do this task with the help of the system.
	Map T1-T5 to text templates	Knowing T1-T5					Only the computer can perform this task
	Structure interface based on generated templates	Modify interface					Only the computer can perform this task
Show written explanations via templates	Navigate Interface	Access to interface					This task needs either party and thus required OPD
	Generate experiment data set	Knowing whole data set					Only the computer can perform this task
	Display data from experiment data set	Modify interface					Only the computer can perform this task
	Predict classification result	Understand ML algorithm					While the computer can do this task, it is targeted towards the user which needs assistance. Thus requiring observability and directability
Human predication of system result experiment	Show correctness of prediction	Know ground truth					Only the computer can perform this task

Legend	
Performer	Support
I can do it all	My assistance could improve efficiency
I can do it all but my reliability is > 100&	my assistance could improve reliability
I can contribute but need assistance	My assistance is required
I cannot do it	I cannot provide assistance

Interactive

Data Browsing: The ability to browse the training data of a machine learning model was universally requested by the interviewed professionals. The browsing ability is extended by different filters for the specific functionality. The user shall be able to explore the training data based on the class of the image. Furthermore the data will be algorithmically grouped based on similarities to potentially expose patterns. Additionally the user shall be able to explore data points that belong to the edge cases of the models capabilities, such as false negative and false positives.

Experiment: Comparing the machine learning models predictions to the human expertise also was a key aspect of AI interaction in the medical domain as shown by the interviews. Therefore a input-output-experiment shall be implemented, where the user is able to test the AI against its own expertise.

Visual

Occlusion: A perturbation based approach to compute attribution, involving replacing each contiguous rectangular region with a given baseline, and computing the difference in output. Occlusion is most useful in cases such as images, where pixels in a contiguous rectangular region are likely to be highly correlated (Facebook, 2021; Zeiler & Fergus, 2013).

Anchors: The algorithm provides model-agnostic and human interpretable explanations suitable for classification models applied to images, text and tabular data. The idea behind anchors is to explain the behaviour of complex models with high-precision rules called anchors. These anchors are locally sufficient conditions to ensure a certain prediction with a high degree of confidence (Ribeiro et al., 2018).

LRP: The Layer-wise Relevance Propagation (LRP) algorithm explains a classifier's prediction specific to a given data point by attributing (positive & negative) relevance scores to important components of the input by using the topology of the learned model itself (Lapuschkin et al., 2019).

Referencing the chapter 2, the main goal of a visual explanation shall be the validation of the AI model's behavior. Therefore a simple technique, such as Occlusion, which only

highlights image regions with a high importance to a result, is reasonable to implement. While LRP does have certain benefits because of its ability to assess positive and negative influences of pixels to the model output, it also comes with a high complexity that manifests as a lot of hyperparameters, making it hard to implement and use properly. Anchors work differently to the other two methods, since it is able to find a hyperpixel that is most influential to a AI's classification. In the context of standardized medical images, that are already preprocessed in a certain way (e.g. X-ray images), the method looses it's edge. The high similarity of images in medical binary classification tasks lessens the expressiveness of Anchors significantly.

3.4 System Architecture

The goal is to create an interactive software application, therefore a suitable system architecture has to be constructed to suite the needs of the users and the usage context. Although, implementing the functionalities from section 3.3 is possible in many different ways. It is possible to realize the assessment system as a classical, offline software application or by leveraging web based tool sets for a possibly distributed cloud solution. Also it is conceivable to implement the system on a middle ground of those two, by creating a software that is build to be ran locally in common web browsers. To stay hardware and software agnostic these three options will be shortly evaluated against each other based on the requirements set by the functionalities.

The common requirements are the ability to store machine learning models and training data for the assessment to be computed. In addition the user has to be presented with a graphical user interface (**GUI**) to interact with. Having these two components in a close relation can drastically reduce the overhead of implementing the communication between the computational and data storage component with the GUI component. On the other hand such a close relation in an offline system can significantly reduce the flexibility of the implementation regarding GUI and interaction design. Furthermore a single offline application has to take many different execution environments (operating systems) into account, which might be a big downside depending on the actual context of usage. Separating the system into a multi-tier application allows for more modularity and freedom in choosing the actual implementation technology. A multi-tier web application allows for a very specialized choice of tools for the respective component at the cost of a higher complexity and implementation cost. Such web applications have the advantage

of being relatively easy to transform into a local application without the need of hosting a server environment. This can be the middle ground between the offline local and the web based distributed application.

Referencing the usage context of the application it is not needed to implement a highly complex distributed application, although Clearbox has shown that it is very much possible to implement a robust cloud based solution. To reduce the scope of implementation for this thesis a middle ground is the most reasonable: Leveraging modern web based tools that are mostly environment agnostic to build a flexible application that could be ran in either the cloud or locally on a single machine. Figure 3.1 shows a possible system architecture for a two-tiered application, where the GUI is separated from the logic and the data store. Such a separation of concerns on the macro level enables the usage of specialized tools for each component. While it is conceivable to move the data store into a separate tier (making it a three-tiered architecture), there are no concrete requirements for using a individual data store technology, such as a dedicated database.

The *frontend* component encompasses the user interface, which will probably be realized with web-based tools as mentioned before. This web-based component facilitates the flexibility of the implementation, as the user will remain mostly hardware and software independent by leveraging common browser technologies. The *backend* will be decoupled from the GUI and therefore can benefit from other technological stacks, optimized for the tasks of machine learning and data science. Additionally the backend will envelop the data and the AI model itself, for providing its services to the frontend. This split allows for a distributed, hardware and software agnostic architecture which can be run either locally or remotely, whereas leveraging the optimal tools for each task.

The communication between the two components will be realized through an **API**. The protocol used for such a communication will be the standardized *Hypertext Transfer Protocol*, which allows for systems to be built independently of the data being transferred (Nielsen et al., 1999).

3.5 Interaction Design

The interaction design is a key aspect of HCI, as it defines how the user will actually interact with the system and how information will be made accessible to the user. Considering the human factors in the design process is very important as described by

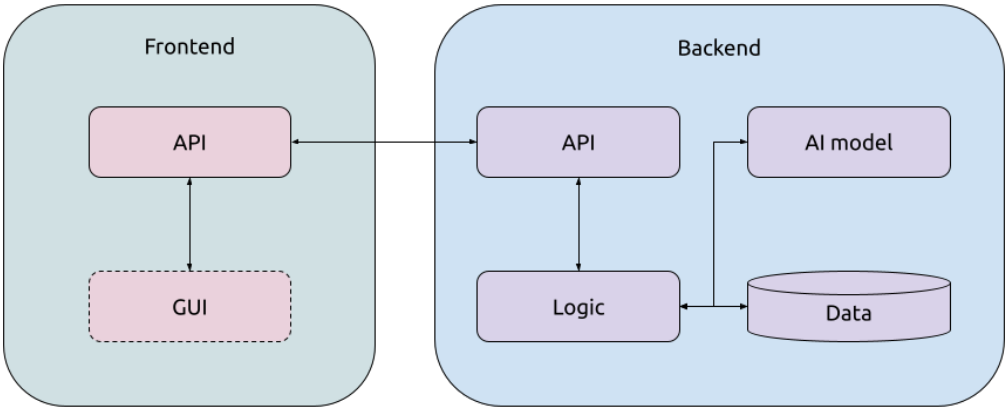


Figure 3.1: Multi Tier System Architecture

Wickens et al. (2016) - for this builds upon the insights of the previous chapters, namely the definition of interaction (sub-)tasks and the interdependency analysis, which in turn are based on the general objectives and user needs as described in chapter 2.

The conception of functionalities and interaction design was the main topic of the previously mentioned expert workshop, which was conducted with colleagues from the University of Lübeck and Clearbox. The goal was to present, validate and discuss the scope of functionalities (based on the previous research and the user requirements) and the interaction design which utilizes scaffolding and guidance for the main user group. The workshop was therefore an iteration of the previously conceived results and the yielded good feedback on the scope of functionalities and the guidance concept. A main focus of the discussion was type of guidance which was then resolved to a semantic signal manifested by an intention query. Another relevant results of the workshop were: (1) The identification of a possible contradiction between different visual explanation techniques, which were described in subsection 3.3.1. This reinforces the idea to limit the visual explanations to one to avoid unnecessary confusion of the user. (2) Feedback from Clearbox's own tests showed the tendency to a more sequentially user experience in contrast to a dashboard centered application. (3) Pre-computation of computationally expensive explanation methods to increase the interaction performance and response time of the application.

A central consideration of the interaction design is to adapt the system to the abilities and needs of the actual users. Referencing subsection 2.1.2, a main user group has low expertise in machine learning, which in turn means special requirements for the

interaction design of an AI assessment system. Therefore the design approach will revolve around a guided interaction version and a unguided interaction version: The guided interaction will leverage a more streamlined approach which facilitates the intentions of the user by asking and then presenting the respective functionality. Trigg (1988) already proposes local guidance of the user by the author of a system, by using intention queries to determine a suitable path through the content of an application. This mode of interaction aims to relieve the user of mental workload by scaffolding the application and thus avoiding the presentation of all functionalities, and therefore aiding the learning process for machine learning topics (Soloway et al., 1994). The other mode of interaction will be unguided, as in there will be no intention query and the user will be able to freely explore the whole lot of functionalities offered by the application, which is hypothesize to be beneficial for users with more expertise in machine learning. The baseline for both ways of interaction is the whole array of functionalities as specified in Table 3.1.

To further explore and concretize the interaction design two methods were applied: *Interaction Dialogues* and *Interaction Flowcharts*. Interaction dialogues aim to creatively explore different ways of implementing the functionalities as defined in section 3.3. Additionally they enable the exploration of other perspectives based on a natural interaction mode (spoken dialogue). The result is then the baseline for the interaction flowcharts, which specify the concrete flow of information in a standardized manner.

3.5.1 Interaction Dialogues

The idea of using interaction dialogues to creatively explore different ways of interacting with the assessment system came from the cooperation with Clearbox. A colleague specialized in building User Interaction and User Experience proposed this method to be used for developing the interaction design. Using dialogues as the mode of interaction and communication between human and computer allows for a very liberated design process. The goal was to explore different variations of interaction based on the functionalities as defined in Table 3.1. An example dialogue (with H for human and C for computer) for task 1 will be presented:

H: I would like to see your training data.
C: Do you want to see all data or just data for a specific class?
H: I want to only see data for class x.
C: Here you go! Do you want to see data of another class, too?

H: Yes please, but give me some time.
C: Of course!

Using the dialogue technique allows for easy exploration of alternative information flows as shown by the next example:

H: I would like to see your training data.
C: Here take a look!
H: Wow, that is really a lot!
C: Do you want to filter for a specific class?
H: Yes, please show me only data for class x.
C: There you go.
H: Thank you, and now please show me all data for class y.
C: Sure, here!

These examples clearly show different ways of interacting in order to accomplish the same task: Browsing the training data. This process was applied to all tasks (and sub-tasks) from Table 3.3 to gather a lot of different interaction variations, to be then used as a baseline for the following interaction flowcharts. The key takeaways from this process, was that there are various possibilities to realize an interaction. With multiple alternatives per task, some comparisons could be made: Often it seems beneficial for a streamlined interaction design, to present the user with a set of options from the beginning, instead of presenting everything and then reducing the amount of information. Furthermore situations with exhaustive searches can be avoided by presenting the user a limited amount of information. Most tasks had two to three alternatives which were compared against each other in order to find the best dialogue, which was then chosen to be the baseline for the standardized flowcharts.

3.5.2 Interaction Flowcharts

Building on the previous chapters the possible human-computer-interactions were formalized into flowcharts. The type of flowchart is defined in *DIN 66001* (Hering, 1984). The flowcharts are essential as a reference for the implementation, as they define the details of the information flow between the two parties. The goal is to have interactions that have a clear start and end point, with no dead ends - leveraging flowcharts for this allows for an easy validation of these goals. Another goal was to separate the applica-

tion into smaller, more manageable pieces, each defined by its own flowchart. Figure 3.2 shows a flowchart that defines the flow of interaction for the process of browsing classified images (task 1 from Table 3.1), with the background of the presented interaction dialogue. For each of the tasks defined in Table 3.1 and Table 3.3 such a flowchart was developed as seen in Appendix C: Interaction Flowcharts.

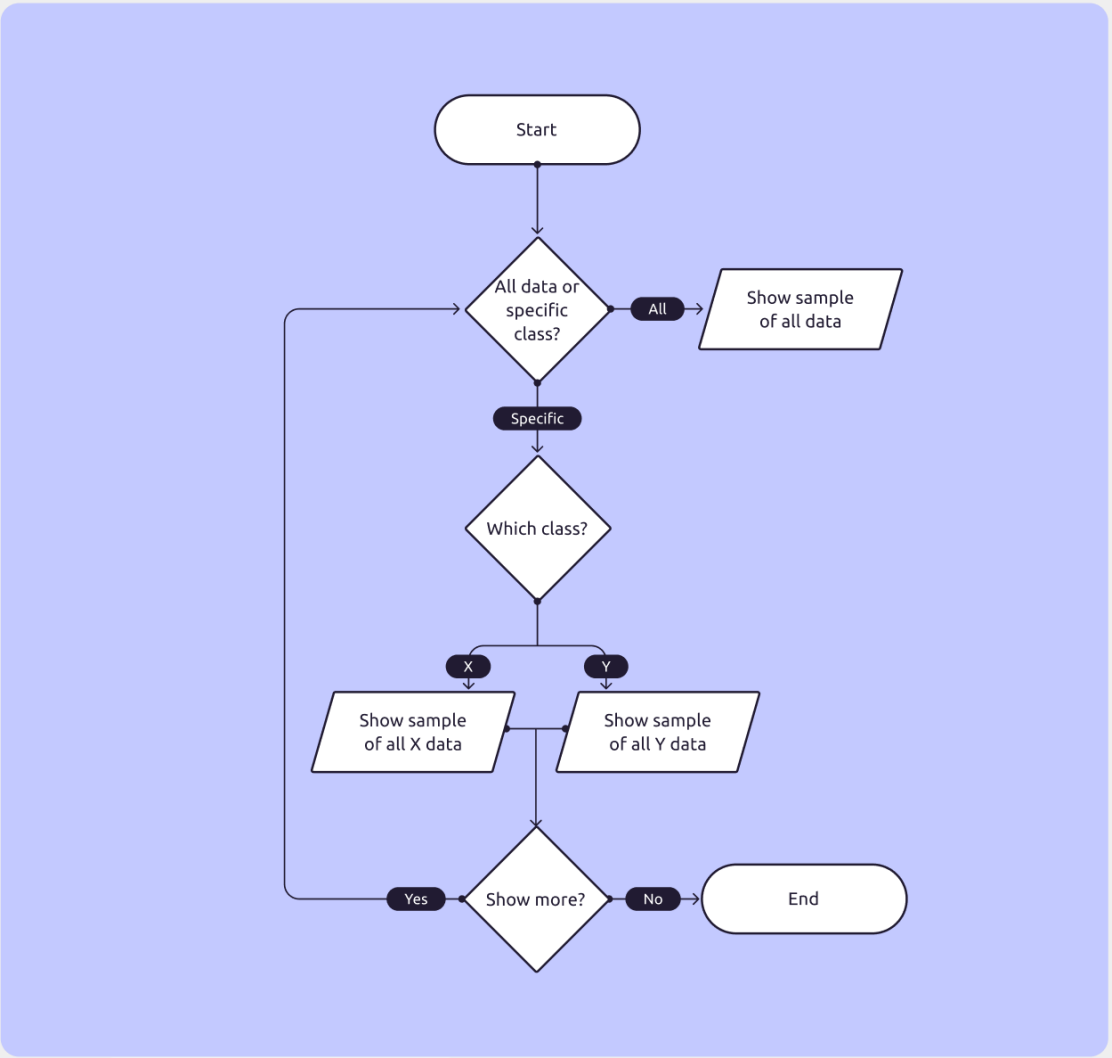


Figure 3.2: Flowchart - Browse Training Data for a given Class

3.6 Interface Design

Building on the functional specification and the interaction design an interface design was developed to encompass the whole array of functionality into a single application. Similar to the flowcharts the application interface will be split up into single elements of interaction as defined by the tasks and interaction flowcharts. This separation allows for a better overview and a more flexible layout. Additionally inspiration was drawn from the state-of-the-art application from Clearbox as seen in section 1.2, where a similar approach was chosen. The concept of splitting the application into smaller pieces also facilitates the usability of the whole application, as the user will not be overburdened with all functionalities simultaneously. Instead the user will be able to choose which functionality he wants to utilize. This layout goes hand in hand with the guidance concept, as every functionality will be represented by its own part of the user interface. This is an important aspect to consider, especially regarding the reusability of software components. Furthermore, preventing multiple implementations of the same functionality helps with the comparability of the guided versus non-guided version of the application.

The application is conceived for devices with relatively big screen sizes, such as desktop computers, laptops or tablets as there was no use case found for a mobile application (see chapter 2). Additionally the screen size is needed for the user to properly view the image content. As such the layout of the application will be vertically scrollable, optimized for landscape orientations to match the devices and the use case.

Figure 3.3, Figure 3.4 and Figure 3.5 show a mockup of the application, designed as a single page layout with all functionalities present. The layout presented here corresponds to the unguided version. On the top of the application (Figure 3.3) general information about the model, its capabilities and standard metrics are shown. Further down (Figure 3.4) the data browsing functionalities are depicted. Lastly the the similarities and the experiment functionalities are shown at the bottom of the application (Figure 3.5). Furthermore the order of the components is determined by the level of involvement needed of the user: The beginning is limited to higher level information, which then transforms into a deeper insight into the training data, model limitations and clustering of data, and concludes in the input-output experiment where the user can test its mental model against the actual AI.

The general design of the application revolves around a simple modern layered design in the form of cards with rounded corners. This aims to make the single components easily

distinguishable while preventing the introduction of unnecessary visual separations. The color theme of the application is chosen to be neutral with high contrast between text and background while utilizing highlighting colors for important interface components for good readability and easy orientation. Although this mockup heavily references Clearbox' design language, a additional dark mode with inverted colors is considered depending on the user preferences and image content.

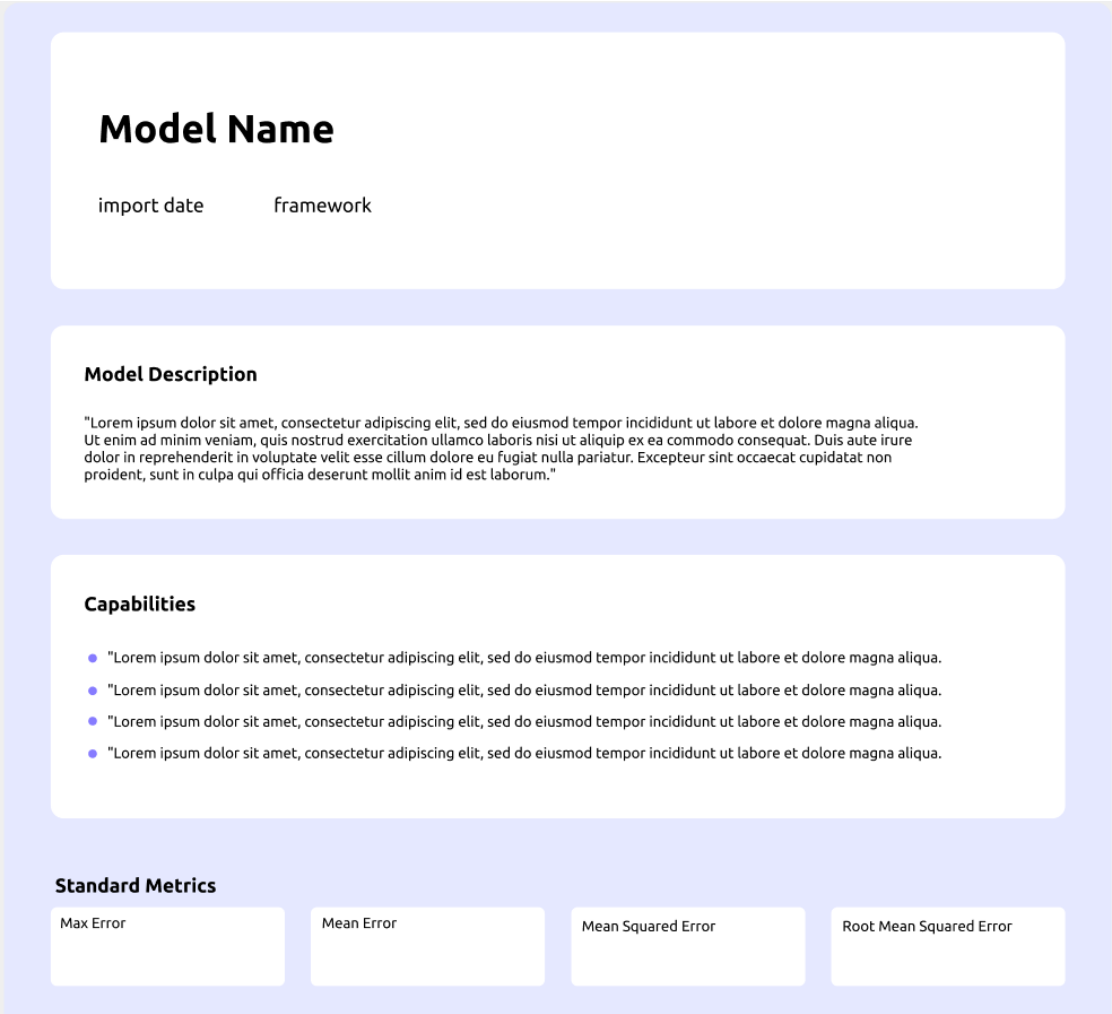


Figure 3.3: Mockup Part 1

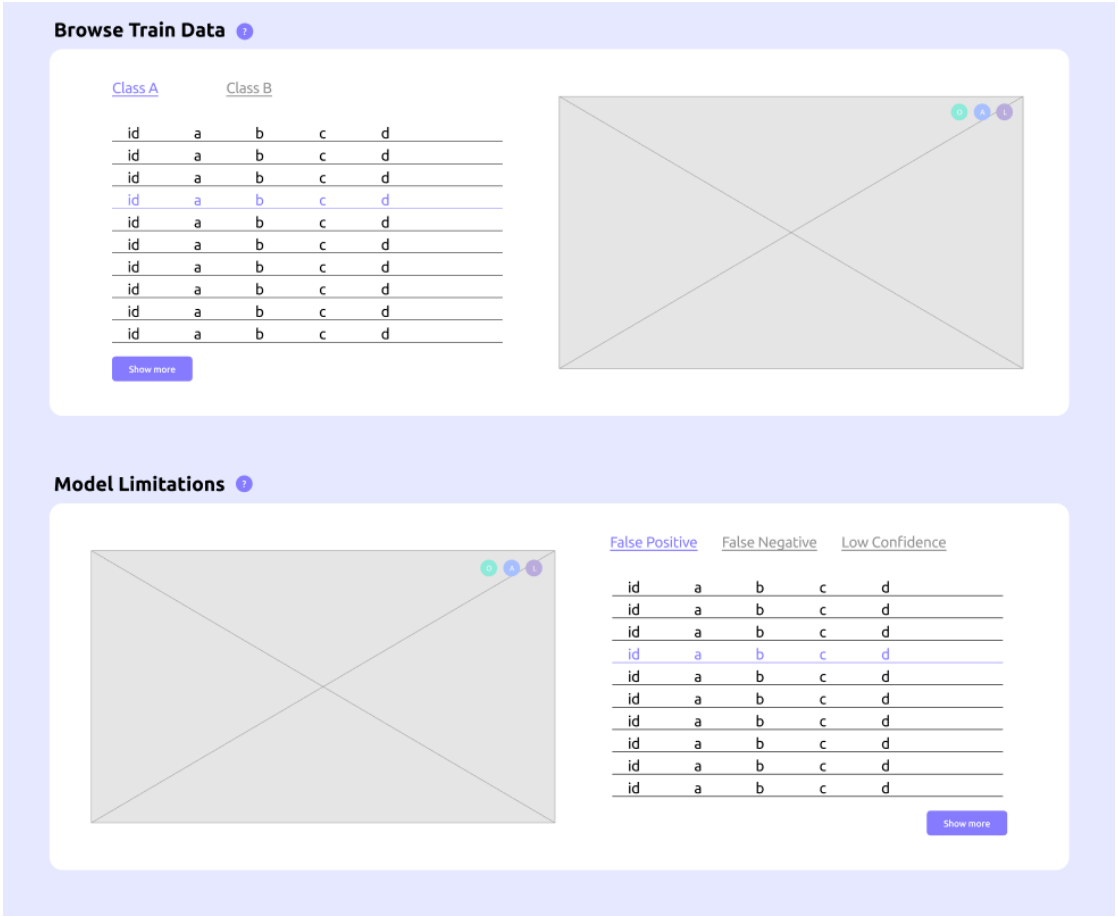


Figure 3.4: Mockup Part 2

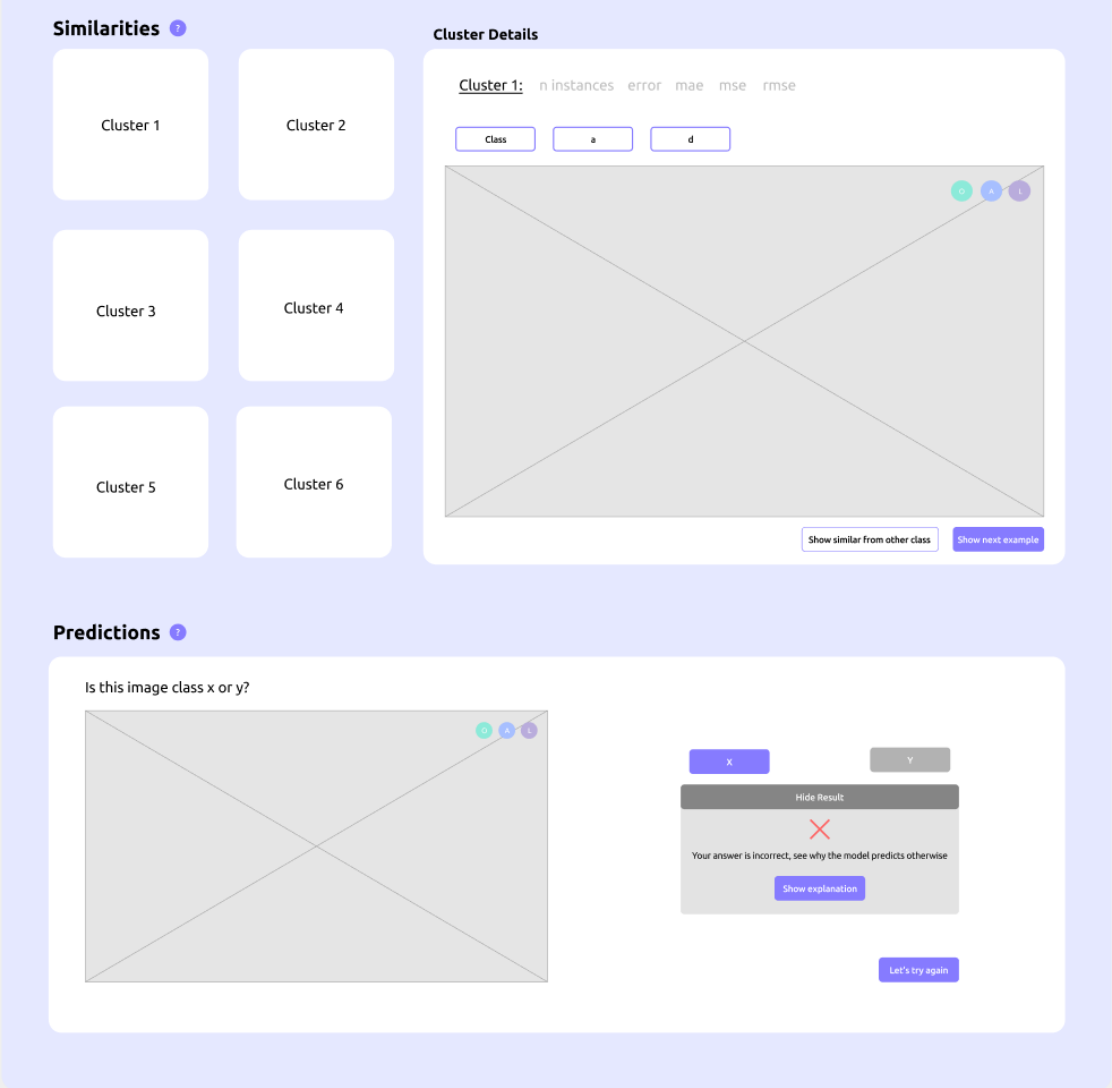


Figure 3.5: Mockup Part 3

3.7 Conclusion on the Conception

Building on the analysis, the conception conceived possible solutions to meet the requirements of the user, the usage context and the overall goals. By referencing the concrete use cases and user requirements seven functionalities were derieved (Table 3.1), which were then analysed for interdependencies in a HCI scenario (Table 3.3), resulting in a clear specification of tasks, sub tasks and capabilities required to perform those for either party of the interaction. Furthermore The functionalities were described in detail in subsection 3.3.1 and insights were collected about the mapping of the functionalities to the explanation techniques, while providing a semantic grouping for those techniques. Although many explanations techniques have a direct mapping, the visual explanation was overrepresented and therefore one from three options was chosen. Having the functional specification set, allowed for the development of a system architecture (section 3.6), which is constructed with flexibility in mind and leveraging modern web based technologies while still meeting the requirements of an AI assessment system. Through a conceptional iteration via an expert workshop the ideas were further discussed and deepened, resulting in interaction dialogues (subsection 3.5.1) for the seven tasks and subsequently in interaction flowcharts (subsection 3.5.2) defining the concrete flow of information between the interacting parties. Finally those functional components were manifested in an interface design (section 3.6), which leverages the interaction design and the idea of guided versus non guided interaction styles to provide a easy to understand GUI for users with low expertise in machine learning. While taking clear reference to Clearbox's design concept and language, the concept here aims to provide two different GUIs that build on the same flexible components, each encompassing one functionality, to be later evaluated against each other.

4 Implementation

The AI assessment system will be implemented based on the insights and conclusions from chapter 3. The system to be build has to realise many tasks from the data science and machine learning domains, while it also needs to provide a GUI for user interaction, as described by section 3.4: Implementing the AI assessment system prototype as a flexible, hardware agnostic application is a main goal of the system architecture and implementation, while also meeting the interaction and design requirements as described in section 3.5 and section 3.6.

Taking these aspects into account, a *Python* based backend implementation is preferable, as the programming language is widely used for data science and machine learning tasks, while also providing tools for web application building. Furthermore Python ranks the most popular programming language as of november 2021 and provides some of the most used and curated software libraries for data science and machine learning among all alternatives (NumPy, 2021; Pandas, 2021; PyTorch, 2021; TIOBE Software BV, 2021). The flexibility provided by Python allows the backend to be completely implemented in said language and making no compromises on the tools needed.

Being decoupled from the machine learning domain, the possible frontend implementation tool set is much more diverse: Popular frameworks for browser based GUIs are *React*, *Angular* and *Vue* (amongst others), all providing the required functionalities in either *JavaScript* or *TypeScript* in combination with the classic *HTML* and *CSS* technologies.

Referencing subsection 2.1.3 and Clearbox, a Python based application leveraging *Streamlit* will be used for implementing the AI assessment system (Streamlit Inc., 2021a). While there are many different possible alternatives to implement such a flexible distributed system (for example *Django* or *Flask* in combination with *React*), Streamlit comes with a big advantage: The ability to directly transform Python scripts into deployment ready web applications including a GUI. This ability completely invalidates the disadvantage

of building a flexible web based application, as it avoids the additional effort of implementing the communication between the presentational and the logic layer of a multi-tier application. Although strictly speaking a Streamlit-based application will not be multi-tiered in its implementation as it only consists of singular python scripts leveraging the Streamlit platform, which in turn hides most of the complexity of implementing a distributed, web based application. Additionally some flexibility in implementing the actual frontend of the application is lost by using such a omnipotent library, as there is no need for a specialized GUI technology. However it is reasonable to limit the implementation complexity of an AI assessment system prototype in the scope of this thesis by leveraging the Streamlit framework.

Also in this conjuncture, it makes sense to limit the implementation complexity of a AI model to be used in the assessment system. Instead of developing a own model, a pre-trained model was chosen. The model used for the implementation of the assessment system prototype is the `densenet121-res224-rsna` model from Cohen, Viviano, et al. (2021). The model is part of an open source software library called *TorchXRayVision* for working with chest X-ray data sets and deep learning models. It provides a common interface and common pre-processing chain for a wide set of publicly available chest X-ray data sets. This concrete model was trained to classify X-ray images and therefore detect a pneumonia disease. Additionally the fitting and also publicly available *RSNA Pneumonia Detection Challenge* data set was used (Radiological Society of North America, 2021). Leveraging a pre-trained model and a curated data set is an important aspect in the implementation of an AI assessment system, as it perfectly resembles an application scenario for the medical domain and therefore supports the implementation of the actual assessment system in the context of this thesis. However, for a complete assessment system implementation, a functionality for importing any machine learning model and data set would be needed - this was omitted, as it is not essential for the evaluation of AI explanation method effects on users.

The whole source code of the AI assessment system prototype can be found digitally on DVD in Appendix A: DVD Contents or online on Gitlab. The following sections will reference parts of the source code when needed.

make repo available

4.1 System Architecture Implementation

Using the Streamlit platform implies some changes to the originally conceived system architecture (see Figure 3.1). The Streamlit platform allows building a whole web-based application with just Python code, and therefore eliminates the need to implement a separate frontend and the communication between the frontend and the backend. Based on the platform's focus on data science tasks, all functionalities (Table 3.1) can be implemented with the provided GUI elements. Figure 4.1 showcases the adapted system architecture, which leverages the Streamlit platform: The frontend shrunk to a thin client, which runs in the browser. The frontend consequently only consumes the service provided by the backend and is responsible for displaying the GUI elements to the user. Furthermore the actual implementation of the GUI elements is already provided by Streamlit: Based on the Python scripts in the backend, GUI elements are generated by Streamlit for the browser to display (see section 4.2 for details). The backend is now embedded in the Streamlit platform and makes use of its API to generate a multi-tier web application by using the provided tools. The internal structure of the backend has not changed and still includes a logic component, the AI model and a data store. The communication between the frontend and the backend is realized via the HTTP protocol as described in section 3.4.

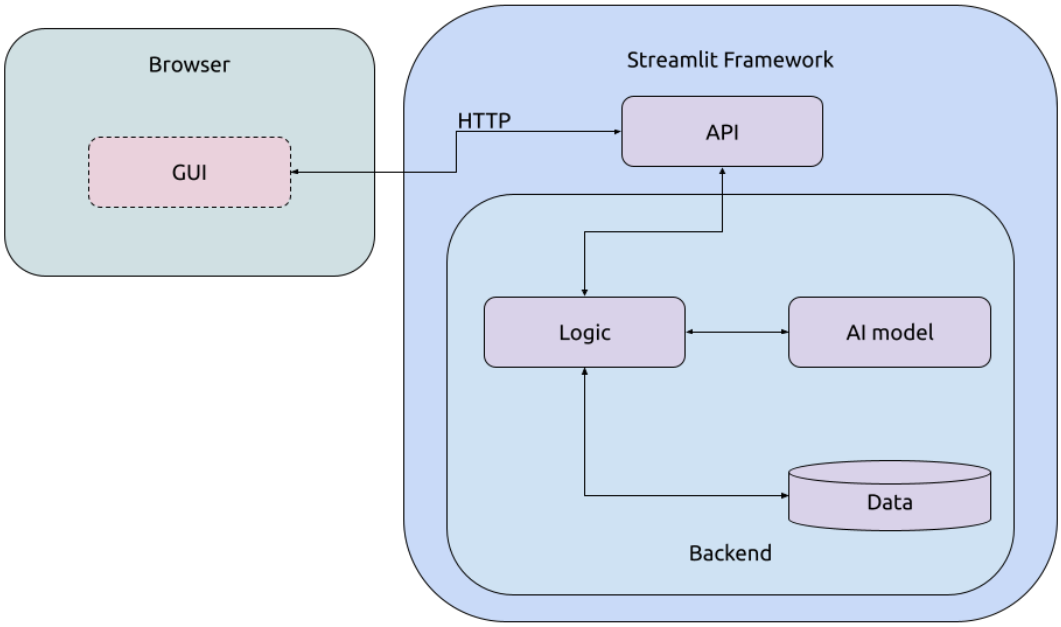


Figure 4.1: System Architecture Implementation with Streamlit

The implementation was split up into two modules, the logic component which handles the user interaction and uses the Streamlit API and the data component which outsources some common functionality regarding data acquisition, transformation and provision. For a basic separation of concerns, these modules were split into two files, where the logic component makes use of the data component, which in turn makes use of the *scikit-learn* and *TorchXRayVision* libraries. Because of the heavy usage of the Streamlit platform, the total system architecture implementation is simple in comparison to a complete multi-tiered solution built on separate technological stacks.

Listing 4.1 and Listing 4.2 show the utilization of the *TorchXRayVision* library and the interaction of the logic module with the data module to load the AI model and RSNA data set for the assessment system to use. Take note of the `@st.cache` annotation, which makes use of the Streamlit API to identify data that can be cached for improved performance. This is especially useful for CSV data, that is readily displayed in the GUI. Listing 4.1 and Listing 4.2 are the foundation for the backend component, which builds on top of the AI model, loaded data and Streamlit API. Table 4.1 shows the primary software used for the implementation of the whole system.

```
model_specifier = 'densenet121-res224-rsna'
model = xrv.models.DenseNet(weights=model_specifier)
d_rsna = load_rsna_dataset()
detailed_class_info = load_detailed_rsna_class_info()
classes = detailed_class_info['class'].unique()
```

Listing 4.1: Initialization of the Model and Data Set

software	version
Python	3.9.6
captum	0.4.1
numpy	1.21.2
pandas	1.3.3
scikit-learn	1.0.1
streamlit	1.2.0
torch	1.9.1
torchxrayvision	0.0.32

Table 4.1: Used Software and Versions

```python
@st.cache
def load_rsna_dataset():
    d_rsna = xrv.datasets.RSNA_Pneumonia_Dataset(
        imgpath='./data/kaggle-pneumonia-jpg/stage_2_train_images_jpg',
        views=["PA", "AP"],
        unique_patients=True,
        transform=transform)
    return d_rsna

@st.cache
def load_detailed_rsna_class_info():
    return
    ↪ pd.read_csv('./data/kaggle-pneumonia-jpg/stage_2_detailed_class_info.csv')
```

Listing 4.2: Functions for Image and Meta Data Acquisition

4.2 Interface Implementation

Based on the concepts of section 3.5 and section 3.6 the interface was implemented using the tools and GUI elements of the Streamlit platform (Streamlit Inc., 2021b). The mainly used GUI elements were: `expanders`, `columns`, `tables`, `images`, `buttons`, `checkboxes` and `selectboxes` - all of which are readily provided. To aid the visual impression of the system, specifically for the display of Xray images, a dark interface mode was made the default, while the user still had the option to change to a bright mode.

As conceived in the conception, all functionalities were implemented as differentiated GUI components by using the `expander` element. Listing 4.3 showcases the implementation of such an component, which makes use of the `expander` element in combination with the `columns` horizontal layout functionality. Specifying the GUI elements with Streamlit builds upon a very structured and sequential convention: Items are presented in the same order, as they are declared in the Python code (with the exception of column layouts). There are very few possibilities to specify complex layout concepts, which automatically results in a very clean and structured GUI. Figure 4.2 shows the rendered GUI element which was specified in Listing 4.3. The special feature of an `expander` element allows for easy hiding or showing of the encapsulated functionality. Figure 4.3 and Listing 4.4 show this exact feature: Multiple functionalities are declared successively with the `expander` element, but only one item is chosen to be presented by the user by clicking on it.

Managing the state of the application is an important aspect, as it allows for the individualization of the user experience. This is particularly relevant for the implementation of the intention query as conceived in section 3.5. The GUI components provided by Streamlit are inherently stateful, data-driven and have a simple life cycle that refreshes on every interaction, which leads to a simple development process that is backed by the underlying data. However it is more complicated to manage state, that is not tied to specific elements and therefore exceeds the life cycle of the element. An example for this is the random choice of image samples to be displayed to the user: The random data points are sampled for each user of the application and shall only be regenerated if the user whishes to do so (see Figure 8.1). Because of the simple life cycle of Streamlit GUI elements, which resets on every interaction, a separate state for some GUI elements has to be managed as seen in Listing 4.5. Streamlit provides a functionality to persist session state per user, which is then saved on the frontend side - this allows a interactive, stateful user experience which is backed by a common data store.

Another important aspect of the interface implementation is the use of the AI model and visual explanation techniques as conceived in subsection 3.3.1. To implement the `densenet121-res224-rsna` model the *torch* library was used. Loading the AI model with this library allows for direct interaction with it, for example the input-output experiment as described in section 3.3. Furthermore the usage of the *captum* library allowed for the computation of visual explanations (attribution by occlusion) for the data set at hand. To increase the performance of the GUI, the explanations were generated beforehand and saved separately to the original image data including the same unique identifier. Listing 4.6 shows the computation of occlusion samples, where each sample results in an image similar to Figure 4.4. The computed images reveal high value superpixels that indicate a high relevance for the AI model. These samples are then used alongside the original Xray images in the assessment system, where the user can choose to display the context-bound visual explanation.

```
1 with st.expander('Overview'):
2     st.subheader(f'{model_specifier}'.upper())
3     overview_l, overview_r = st.columns(2)
4     overview_l.text(f'Import Date: {datetime.date.today()}')
5     overview_r.text('Framework: Pytorch')
```

Listing 4.3: Overview GUI Element

```python
with st.expander('Overview'):
    # [...]

with st.expander('Model Description'):
    # [...]

with st.expander('Capabilities'):
    # [...]

with st.expander('Standard Metrics'):
    st.subheader('Standard Metrics')
    metrics1, metrics2, metrics3, metrics4 = st.columns(4)
    metrics1.metric(label='Accuracy', value=str(metrics['accuracy'].round(2)))
    metrics2.metric(label='Precision',
    ↪  value=str(metrics['precision'].round(2)))
    metrics3.metric(label='Sensitivity', value=str(metrics['recall'].round(2)))
    metrics4.metric(label='F1', value=str(metrics['f1'].round(2)))
```

Listing 4.4: Descriptive GUI Elements

Figure 4.2: Overview GUI Element

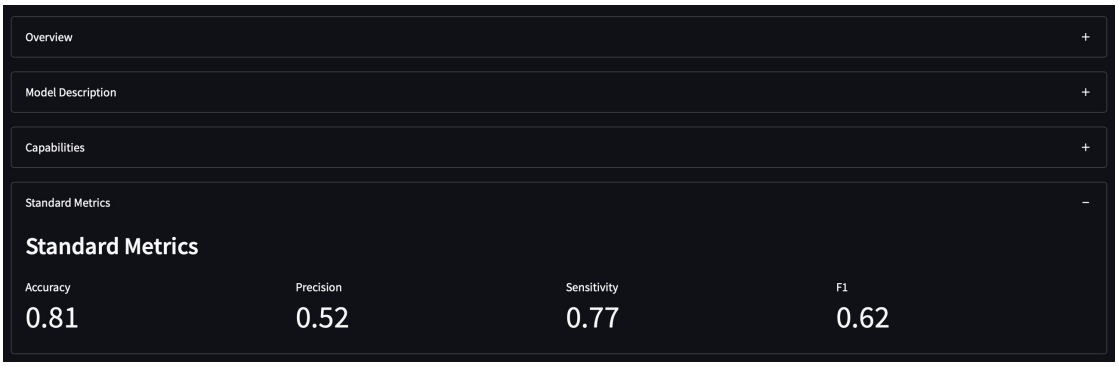


Figure 4.3: Descriptive GUI Elements

```python
def set_browse_indices(high):
    st.session_state.indices = np.random.randint(low=0, high=high, size=10)
# [...]
if 'indices' not in st.session_state:
    set_browse_indices(len(dataset.index))
index_list = st.session_state.indices
df_samples = dataset.loc[index_list]
```

Listing 4.5: Managing State

```python
occlusion = Occlusion(model)

for index, row in df_predictions.iterrows():
    patient_id = row['patientid']

    patient_index = d_rsna.csv[d_rsna.csv['patientid'] ==
    ↪  patient_id].index.values[0]
    print(patient_index)

    model_input = d_rsna[patient_index]['img']
    model_input = np.expand_dims(model_input, axis=0)
    model_input = torch.from_numpy(model_input).float()

    pred_label_idx = model(model_input).argmax()

    attr = occlusion.attribute(model_input,
                               strides=(1, 30, 30),
                               target=pred_label_idx,
                               sliding_window_shapes=(1, 30, 30),
                               baselines=0,
                               show_progress=True
                               )

    plt.imshow(model_input[0, 0, :, :])
    plt.contourf(attr[0, 0, :, :], alpha=0.5)
    plt.colorbar()
    plt.savefig(f'./data/kaggle-pneumonia-jpg/occlusion/{patient_id}.jpg',
    ↪  bbox_inches='tight', dpi=150)
    plt.clf()
```

Listing 4.6: Computing Attribution through Occlusion

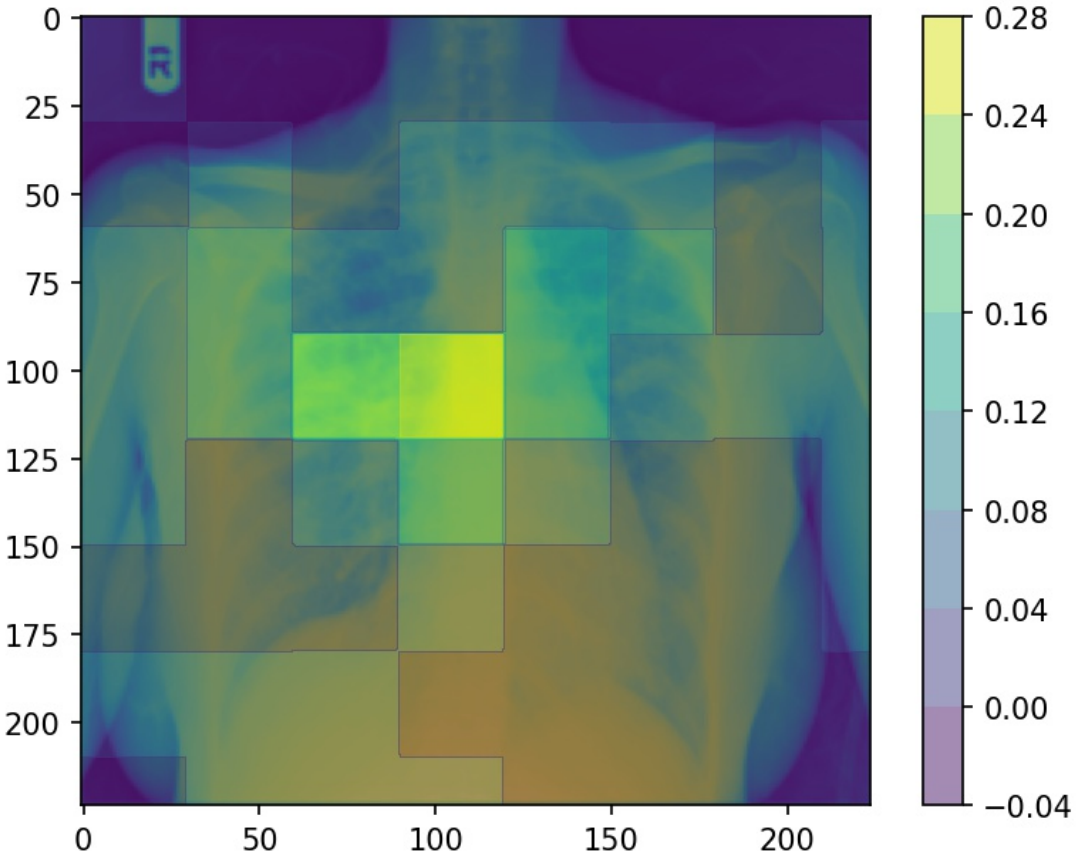


Figure 4.4: Attribution through Occlusion

4.3 Conclusion on the Implementation

Many of the concepts presented in chapter 3 were realized in the AI assessment system prototype, while some aspects grew to be differently than originally conceived, due to technical compromises. The previous chapter showcased the some important aspects of developing a flexible, multi-tier web application with the focus on the functionalities defined formerly. The technologies used for this implementation are mainly Python and the Streamlit framework, which offer a exceptionally well match for the task at hand. By leveraging the data-science oriented framework, a GUI for common web browsers could be realized directly from the Python scripts working directly on the AI model and data. The complexities of developing a distributed web application could be largely avoided by using the Streamlit platform - this also shows in the adjusted system architecture. However some flexibility in implementing the actual GUI design was lost, due to the already provided GUI elements. Instead of developing custom interactive elements with a specialized frontend technology, the prepackaged elements were used. The functionalities from section 3.3 could all be implemented with the exception of functionality number three. The absence of this functionality is caused by the failure to find a suitable algorithm to find commonalities in XRay images to be applied to this data set. Additionally the actual design of the GUI is also predefined by the Streamlit framework, while still maintaining clear similarities to the concept of section 3.6.

5 Dialogue Samples

Based on the main use case (see section 3.2), the realized application will be presented in the following chapter. Leveraging a common use case, allows for an exemplary tour through the AI assessment system. The baseline is the unguided version, which gets supplemented by the guided version. Additionally mobile versions (smartphone & tablet) will be showcased exemplarily. The application was implemented in english (shown here) and German (used for the evaluation). All versions of the application (english / German / guided / unguided) can be found digitally in Appendix A: DVD Contents or the Github repository.

Figure 5.1 shows the starting point for the interaction. Opening the application in the browser leads to this landing page. The page presents the user with all functionalities in the unguided version. Alternatively the guided user gets a slightly different starting point as shown in Figure 5.2, as the guided version hides the functionalities behind the intention query on top. Following the unguided interaction flow, the user is now free to explore all functionalities by clicking on an expander GUI element. A typical starting point for the user are the first three elements as shown in Figure 5.3. Figure 5.4 shows a possible next interaction, where the user decides to browse the AI training data. The figure highlights the focus on class-based filtering and a small sample size, that can be regenerated. The user is able to select a data point and view the image data. Additionally the user can supplement the displayed information with metadata as shown in Figure 5.5. By opening or closing the expander elements, the user has the freedom to shape the vertical layout of the application. Closing the other elements allows for a better overview and navigation as shown in Figure 5.6, while focusing on the current functionality. The input-output experiment may be a interesting next interaction point for the user, because it was commonly requested as described in section 2.4. To further deepen the understanding of the AI, the user might want to see how the AI came to its results. Figure 5.7 shows the visual explanation method (see subsection 3.3.1), which the user can use to gain additional information about the AI's reasoning. Figure 5.8

shows the alternative guided experiment functionality for comparison. Finally the user might want to explore some more detailed information about the data set and therefore go to the data clustering functionality as depicted by Figure 5.9, where the user can gain insights about statistical patterns recognized in the meta data. Although not all functionalities were used, such can be a exemplary user scenario. Furthermore the application is designed to be responsive in regards to the used device. Figure 5.10 shows the previously omitted model limitations functionality on a medium sized device, such as a tablet. Using such a device has the advantage of a possible vertical screen orientation and therefore being able to present more information vertically. Figure 5.11 supports this statement, where the user can interact with the whole experiment functionality in contrast to Figure 5.6. However further reducing the screen size leads to a cluttered interface: Figure 5.12 and Figure 5.13 show the overview, metrics, browse and experiment functionality on a smartphone.

Figure 5.1: Dialogue Sample - All Functionalities (large display)

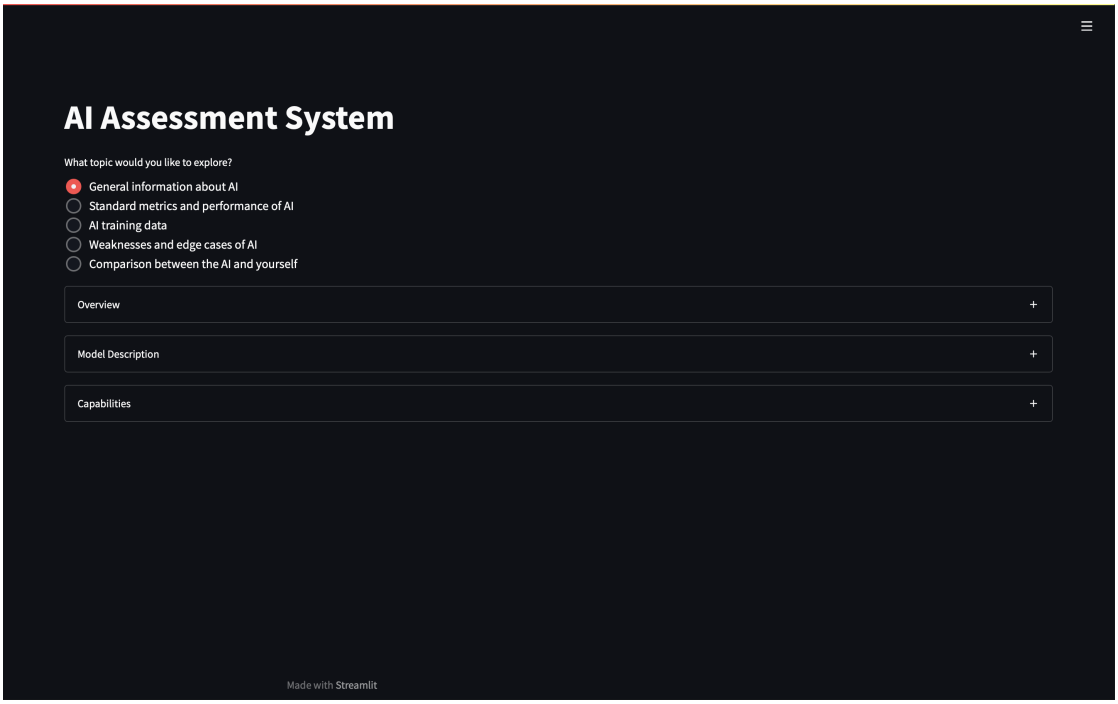


Figure 5.2: Dialogue Sample - Guided Overview Functionalities (large display)

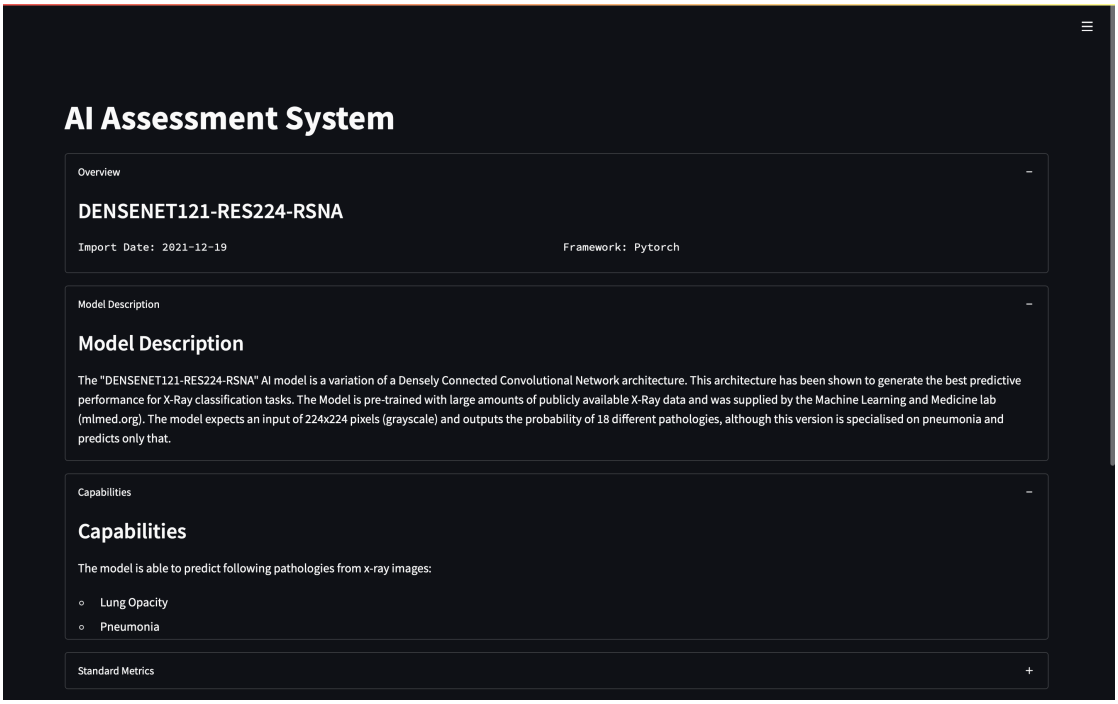


Figure 5.3: Dialogue Sample - Overview Functionalities (large display)

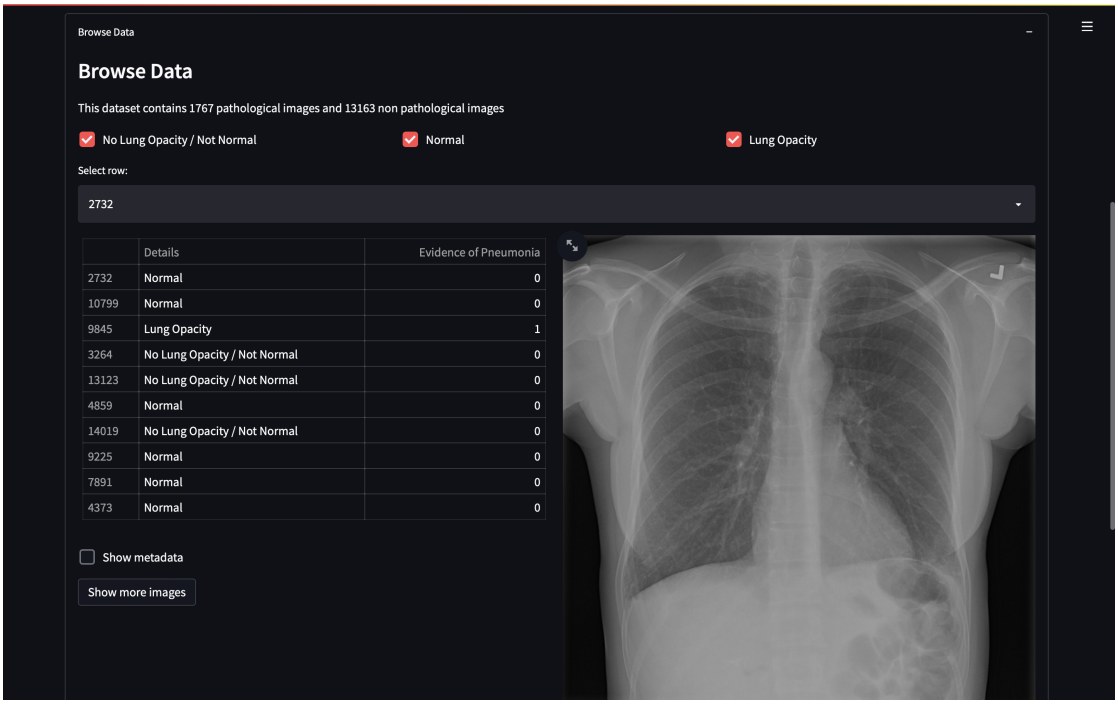


Figure 5.4: Dialogue Sample - Browse Functionality (large display)

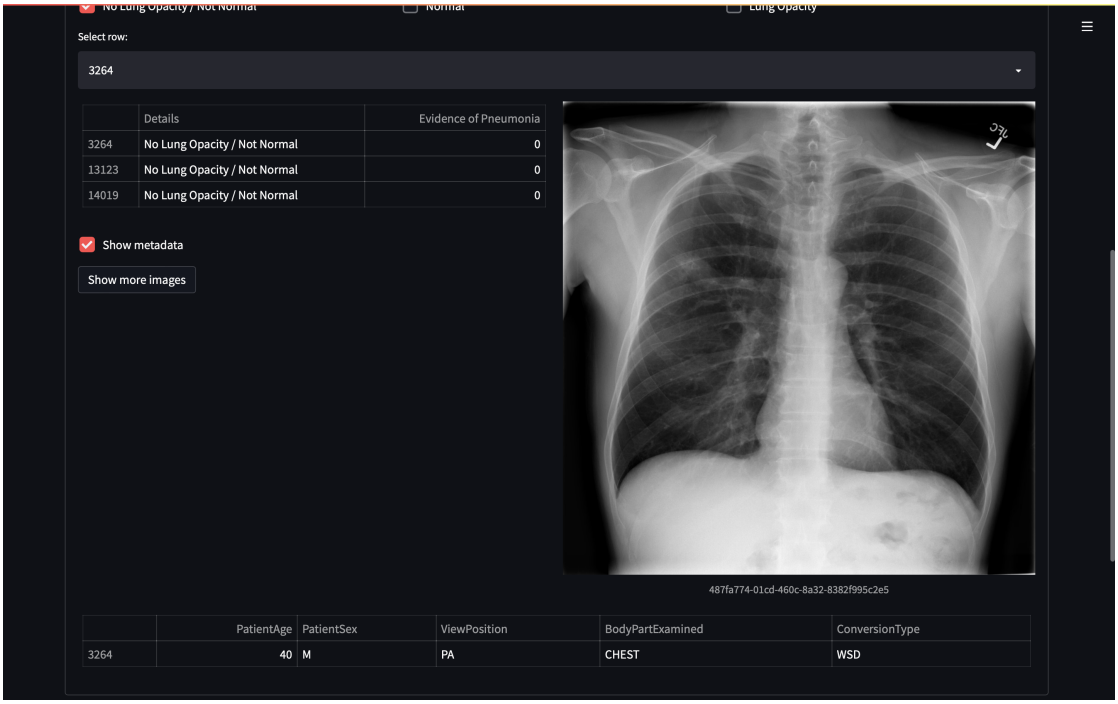


Figure 5.5: Dialogue Sample - Browse Functionality with Metadata (large display)

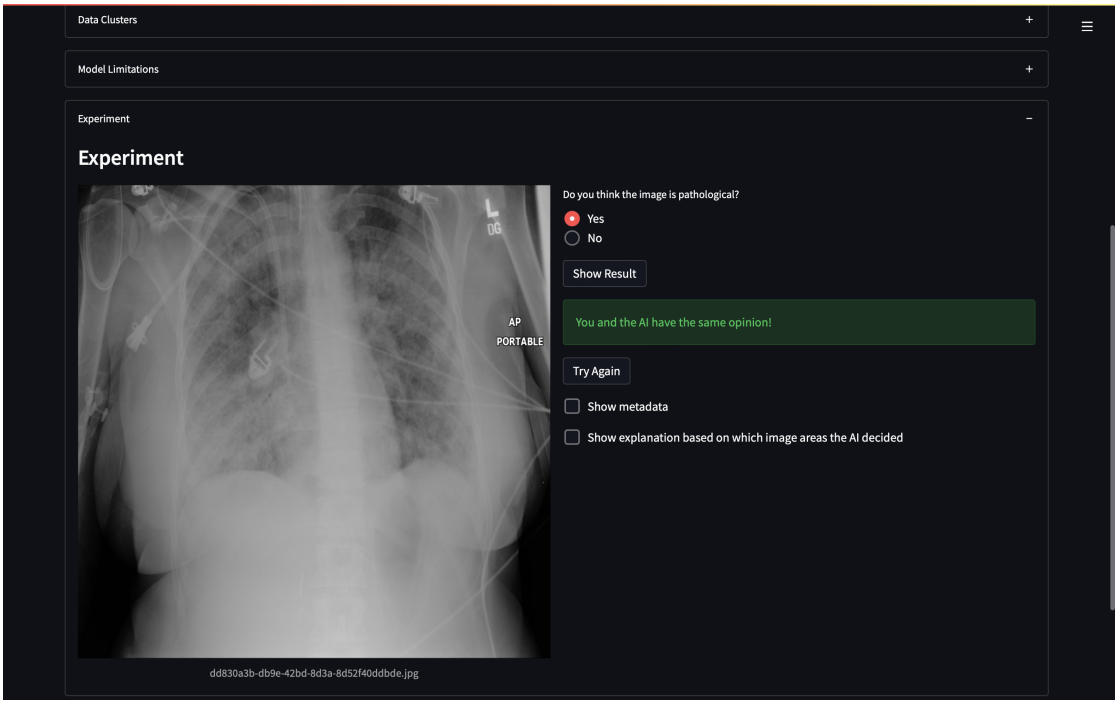


Figure 5.6: Dialogue Sample - Experiment Functionality (large display)

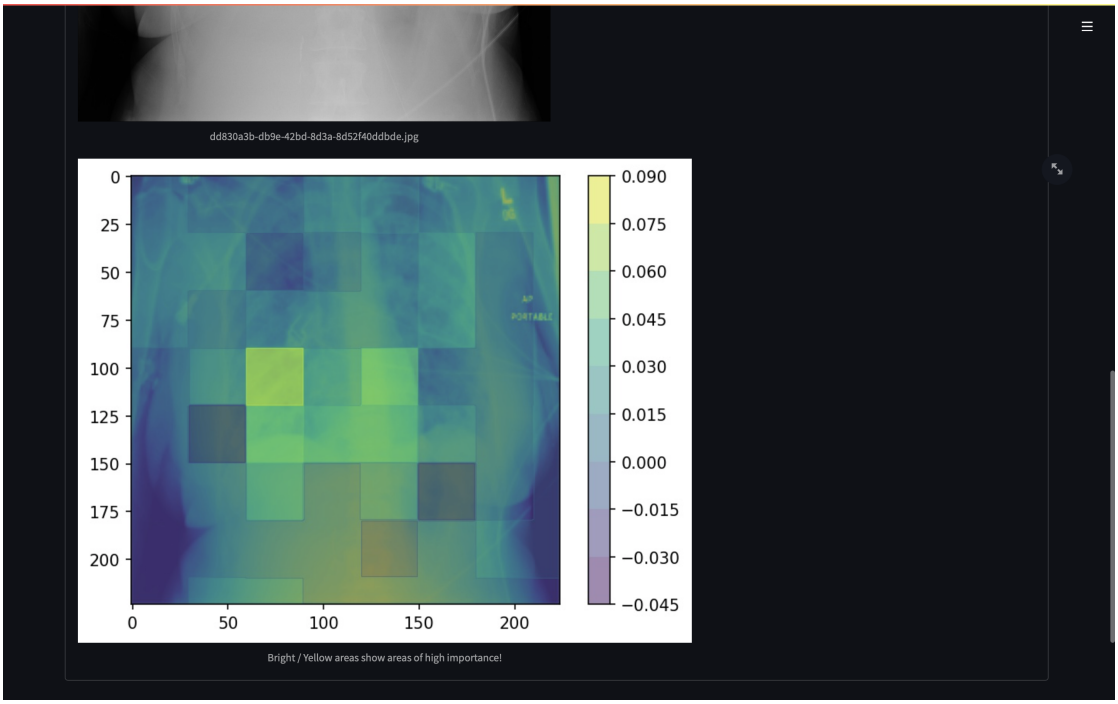


Figure 5.7: Dialogue Sample - Experiment Functionality with Attribution by Occlusion (large display)

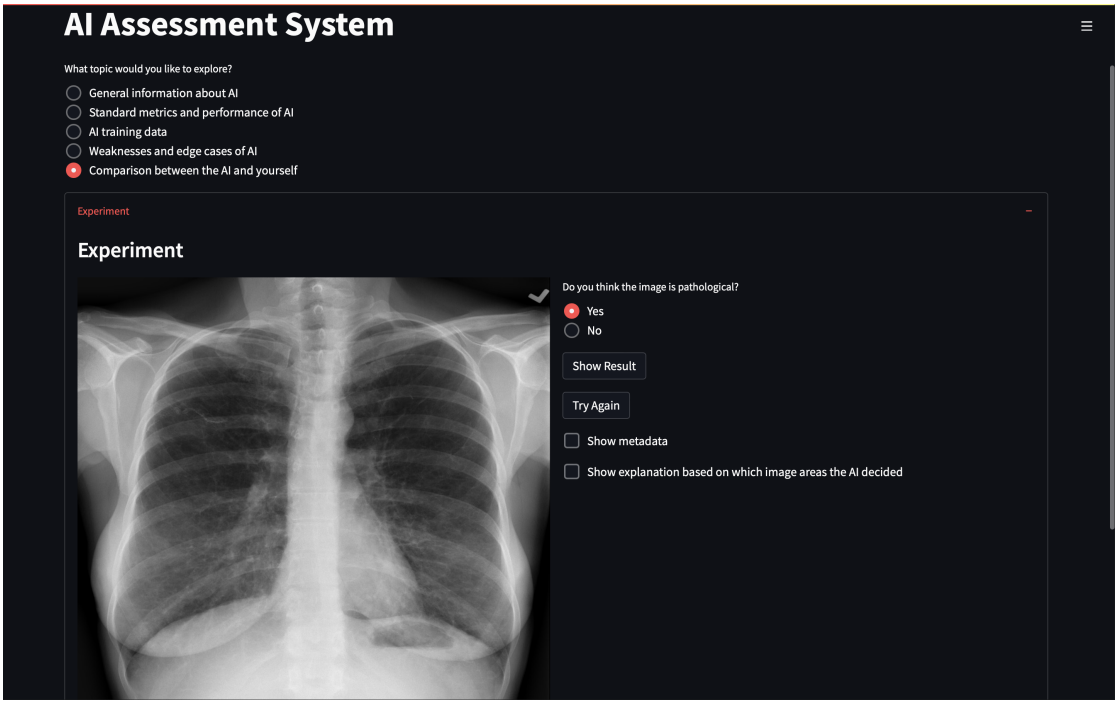


Figure 5.8: Dialogue Sample - Guided Experiment Functionality (large display)

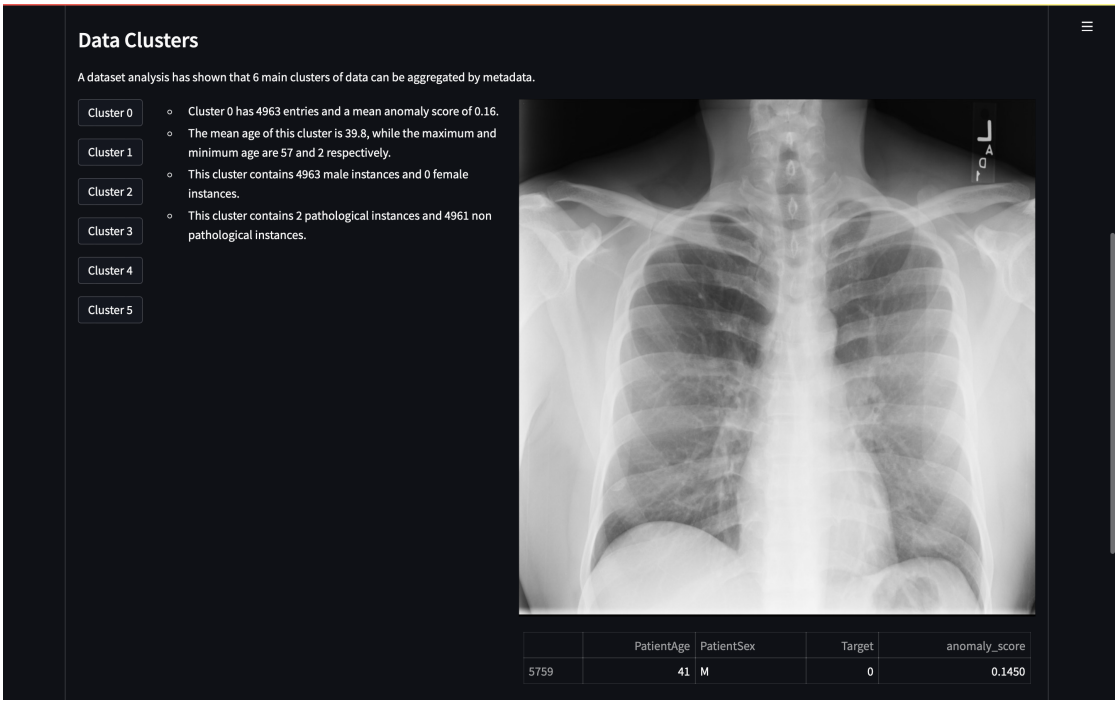
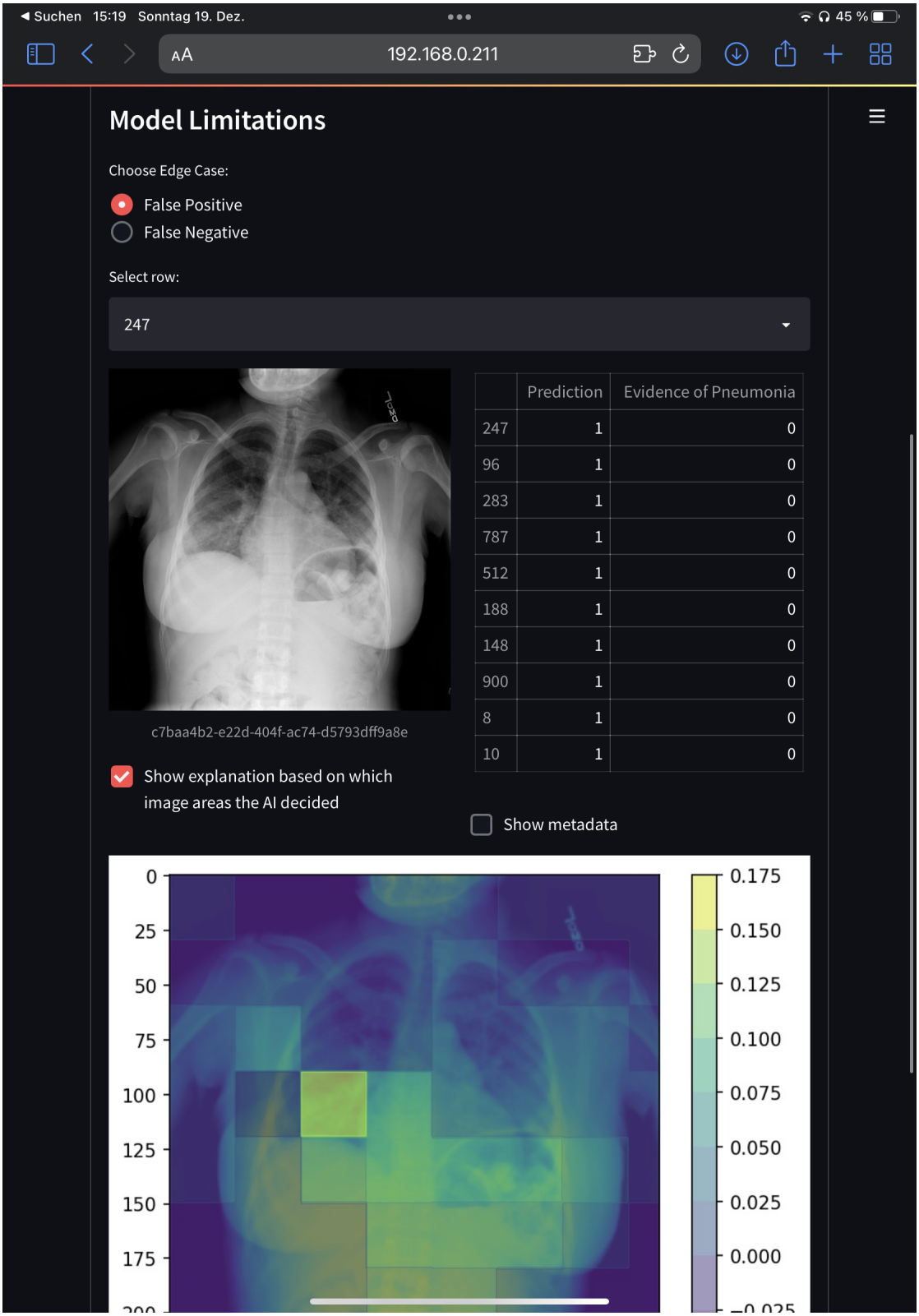


Figure 5.9: Dialogue Sample - Data Clustering Functionality (large display)



	Prediction	Evidence of Pneumonia
247	1	0
96	1	0
283	1	0
787	1	0
512	1	0
188	1	0
148	1	0
900	1	0
8	1	0
10	1	0

Figure 5.10: Dialogue Sample - Data Clustering Functionality (medium display)

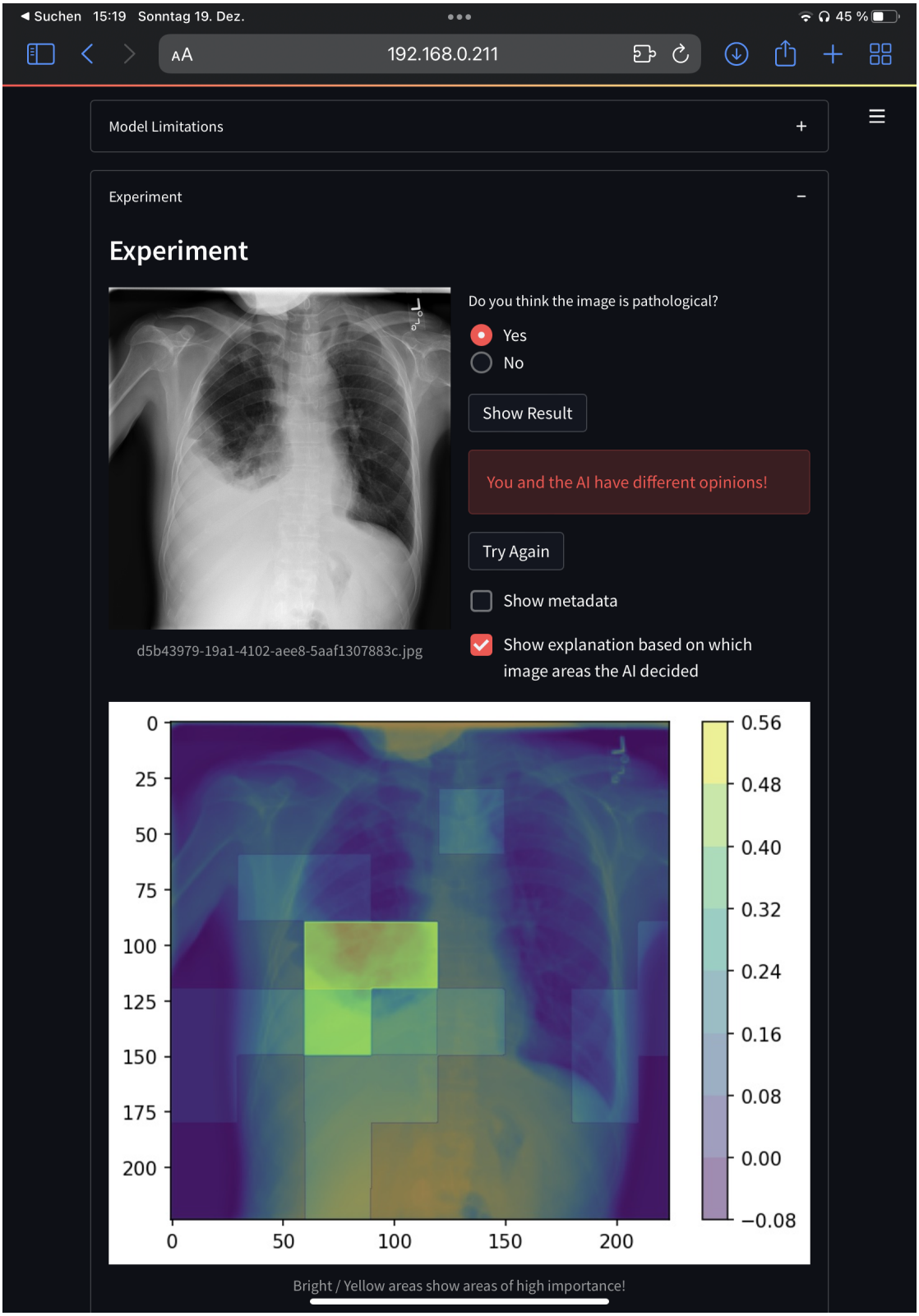
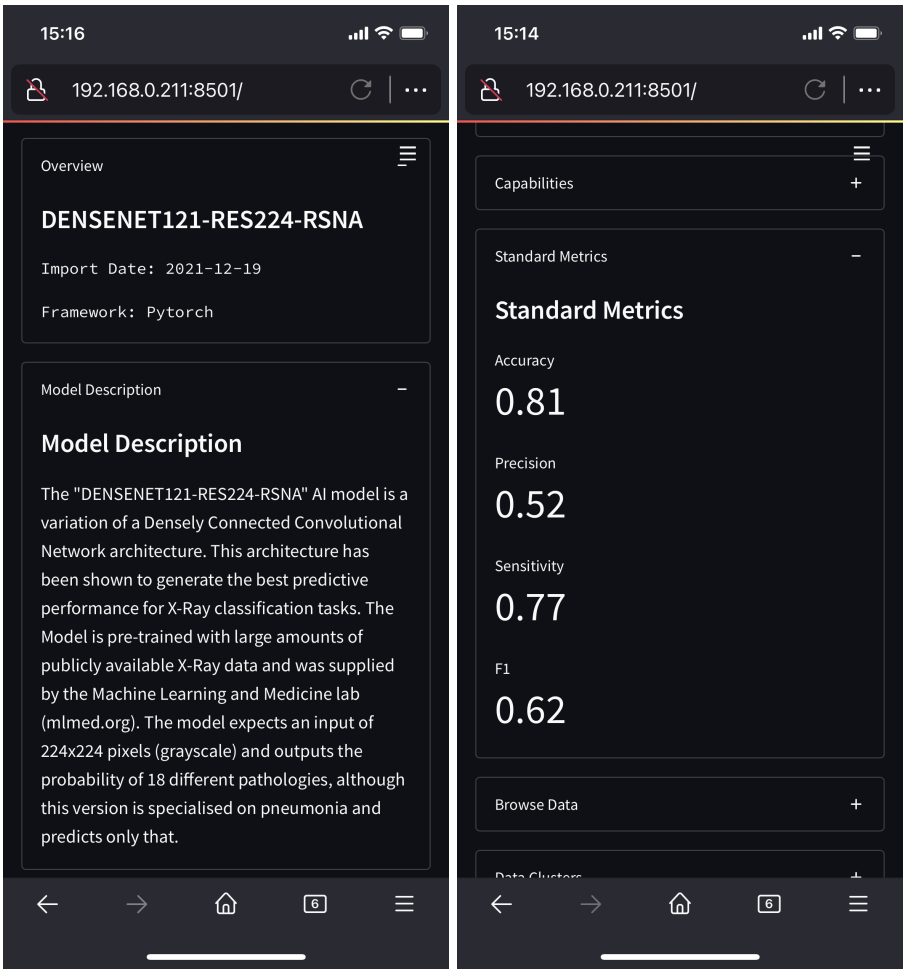
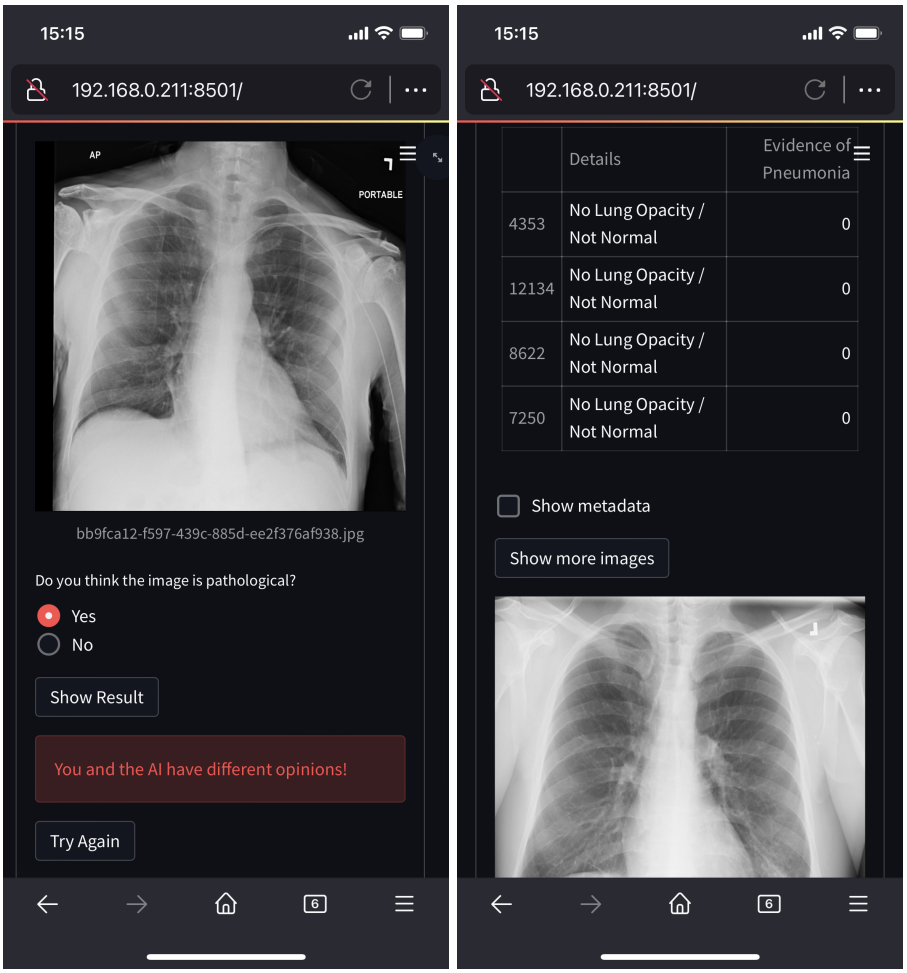


Figure 5.11: Dialogue Sample - Experiment Functionality (medium display)



(a) (b)

Figure 5.12: (a) Overview Functionality (small display) (b) Metrics Functionality (small display)



(a) (b)

Figure 5.13: (a) Experiment Functionality (small display) (b) Browse Functionality (small display)

6 Summative Evaluation

Following the principles of a human centered development of software, a summative evaluation of the AI assessment prototype was conducted. This is important to assess the efficiency, effectivity and quality of the solutions from conception to implementation (Gediga et al., 2002). The following sections will give insights into the concrete goal, the used methods, the study design and the results of the evaluation.

6.1 Goal

In the context of the thesis it is important to study the concrete effects of the AI assessment system prototype on the user regarding information processing, trust and workload. These three aspects are directly derived from the research questions and the goals of the thesis (see section 1.1). Assessing explainability and trustworthiness of AI models through the usage of an assessment system and evaluation with potential users of such system is the main goal. The focus is on the effects of the interactive explanation methods described in subsection 3.3.1. The study also aims to generate insights into the differences of the guided versus unguided interaction modes. The results of the evaluation will create a foundation for further discussion and research on this topic, especially regarding the human factors in the interaction of humans and computers respectively intelligent system, such as AI models.

6.2 Methods

To gain insights on the effects of the AI assessment system on the potential users a on-site experiment was conducted. The study was realized in German, because the place of evaluation was the university of Lübeck, where most participants are only fluid

in the German language. To assess the effects of the AI assessment system and the different interaction modes (guided versus unguided) a mixed study including within-subject and between-subject aspects was developed aiming at approximately 30 subjects. The effects of the independent variables (interaction with the AI assessment system and guided versus unguided interaction) on the dependent variables (Subjective Information Processing Awareness (**SIPA**), Trust and Workload) were studied.

It is important to mention the focus on subjective measures in this study, as the main aspects of trustworthiness and explainability of AI models are also highly subjective. Therefore mainly subjective data was collected - with the exception of one objective workload measure. The complete evaluation questionnaire can be found in Appendix D: Evaluation Questionnaire.

6.2.1 Participants

The participants of the evaluation were chosen to be medical students, as this is largely in line with the main user group described in section 2.4. Although medical professionals such as interviewed in subsection 2.1.2 are preferable but are particularly difficult to make appointments with. Since the study aimed to evaluate approximately 30 participants in a short time, a high turn around count was important. To further increase the initiative of participating in the study a raffle was set up, drawing four winners rewarded with 50€ each.

The study was announced on multiple channels (university internal platform and research partner contacts) two weeks before the three week long execution period. In total, $N = 14$ people participated in the study. All participants were medical students from the university of Lübeck with an age ranging from 21 to 32, 6 of which were male and 8 female. The lowest semester recorded was 5th and the highest was 11th. All participants stated having at least a fundamental knowledge in radiology, but have a very mixed level of affinity for technology interaction. Table 6.1 compiles the information of the 14 participants.

	age	semester	ATI score
N	14	14	14
M	24.36	8.07	3.91
SD	3.34	2.36	0.83
Min	21	5	2.11
Max	32	11	5.33

Table 6.1: Descriptive Statistics of Basic Information on the Evaluation Participants

6.2.2 Design

The study was designed to incorporate within-subject and between-subject components in an one-to-one on-site experiment. The within-subject component is the interaction with the system, since all subjects interacted with the assessment system. The between-subject component is the interaction style (guided versus unguided), as the participants were split into two groups of the same size, one interacting with the guided version while the other interacting with the unguided version. Figure 6.1 showcases the study design with the subject components and the data acquisition points: Firstly, the by e-mail recruited participants were introduced to the context and content of the study. After consenting to a data protection agreement, the participants completed the first questionnaires, which act as the foundation for comparing the information processing and trust before and after the interaction. During the interaction with the system, the time spend on the interaction was measured as a indication of objective workload (with a cut off at 15 minutes). After the interaction the participants completed the last questionnaires. This design allows for comparison of various subjective characteristics before and after the assessment system interaction while also measuring differences between the interaction styles in an overall duration of approximately 30 minutes.

6.2.3 Setting and Instruments

Setting

The experiment was conducted on site at the university of Lübeck. For the timeframe of the evaluation the usability lab was provided by the Institute for Multimedia and Interactive Systems (IMIS), where the study was carried out. The lab was used for the completition of questionnaires and the interaction with the assessment system. The actual interaction was carried out on a 2016 MacBook Pro (i5, 16GB RAM, MacOS

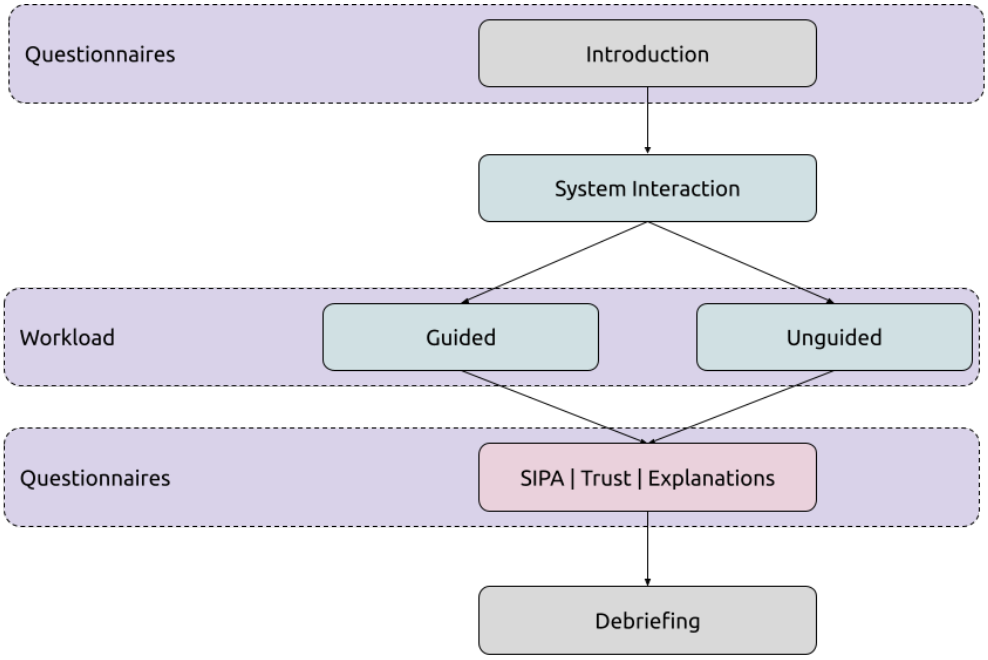


Figure 6.1: Study Design

12.0.1, 13"), where the whole application was running (backend and frontend). The web browser used for displaying the GUI was Safari (version 15.1). Only one participant at a time was evaluated, because special requirements in the form of a hygiene concept had to be met due to the ongoing pandemic situation.

Application usage

The participants were presented with the application as implemented in chapter 4 - either with the guided or the unguided version (see chapter 5). Additionally it was made sure, that the participants could use the peripherals to interact with the GUI in the Safari browser. After the initial introduction, the participants were told to freely interact with the application for a maximum of 15 minutes.

Questionnaires

The study included questionnaires to be answered before and after the interaction with the system. The questionnaires used in this study were:

- Basic Demographic Query (pre)

- Subjective Information Processing Awareness Scale (SIPA) (Schrills et al., 2021) (pre & post)
- Facets of System Trustworthiness (**FOST**) (Franke et al., 2015) (pre & post)
- Affinity for Technology Interaction (ATI) (Franke et al., 2019) (pre)
- Explanation Satisfaction Scale (**ESS**) (Hoffman et al., 2019) (post)
- NASA Task Load Index (**NASA-TLX**) (Hart, 2006) (post)

The whole compiled questionnaire with all sub questionnaires and explanatory texts in German can be found in Appendix D: Evaluation Questionnaire. The questionnaire results were digitalized and aggregated in an excel sheet to include all answers per participant (each participant got a row in the sheet identified by a pseudonym). Per participant the results of each sub-questionnaire were evaluated by calculating the mean with regards to inverted items, as described by the original authors (Franke et al., 2019; Franke et al., 2015; Hart, 2006; Hoffman et al., 2019; Schrills et al., 2021). Each resulting mean value was appended in its own column.

6.2.4 Procedure

The participants were invited individually to the study in the usability lab of the university of Lübeck. Firstly the participants were greeted and the necessary covid-19 measures were taken - only vaccinated, recovered or tested people could participate under strict hygienic requirements. After the greeting, the briefing started. The briefing included a spoken and written thematic introduction and a brief overview of the procedure. Additionally the participants got to read an explanation of the AI model used, the standard metrics and the assessment system. When the briefing was completed without open questions, the participants needed to consent to a data privacy agreement. This marks the start of the data acquisition, which takes place on paper (printed version of Appendix D: Evaluation Questionnaire). The first half of the questionnaire included demographic, ATI, SIPA and FOST. SIPA and FOST here are referring to the explanation of the AI model in the briefing. After completing the first half, the interaction with the assessment system was started. The participants were asked to take place in front of the prepared laptop device with the instructions:

> "Your goal is to gather insights about the model's training data, strengths and limitations. Try to understand how the model is working through the interaction with the assessment system. For this you have 15 minutes. If you think that you will not be able to make any further discoveries before the 15 minutes are up, feel free to stop the interaction. If there are any questions or errors, feel free to ask for help."

During the interaction the time was recorded. After the interaction the participants got to complete the second half of the questionnaire including SIPA, FOST, ESS and NASA-TLX. This marks the end of the data acquisition through the written questionnaire. After the competition, the debriefing started with the participants being asked whether there were things that went not so well, that struck them or created questions. Furthermore the participants were asked if they have comments on things that they liked about the interaction or general comments on the study. The data from the spoken debriefing was written down to complement the quantitative data.

6.3 Results

The descriptive statistics of the results are depicted in Table 6.2 and form the baseline for further analysis. These descriptive statistics can additionally be grouped up by the participant's interaction mode (guided vs. unguided), which is showcased in Figure 6.2, Table 6.3 and Table 6.4. When comparing the SIPA and FOST pretest to the posttest scores, the statistics show that the interaction had effects on the explainability and trust. The objective time measured for the interaction was no indicator of the workload of the participants, as almost all used the full 15 minutes - for this the NASA-TLX score is more insightful as the unguided group has an overall higher task load index. Also the explanation satisfaction is overall higher in the unguided participant group.

To further evaluate the sampled data, a repeated measures analysis of variance (**ANOVA**) was conducted, in which the subjects are measured more than once to determine whether statistically significant change has occurred (Vogt & Johnson, 2011). Because there are two independent variables (interaction & guidance) a two-way ANOVA was applied to the data set from Table 6.2 for the dependent variables SIPA and trust (FOST). Table 6.5, Table 6.6, Table 6.7 and Table 6.8 showcase the results of the ANOVA. The results show a statistically significant change on the SIPA within-subject effect and the

trust within-subject effect. However the between-subject effects do not display a statistically significant change. Although trust increased, SIPA decreased in the posttest compared to the pretest. Also the correlation between SIPA and trust increased through the interaction with the system as depicted in Table 6.10 compared to Table 6.9. Interestingly ATI does not affect any other variable.

As mentioned in subsection 6.2.4 there was also qualitative data recorded after the interaction. Overall the assessment system was well received and most participants found it interesting to get insights into an AI model. The input-output experiment functionality was commonly stated to be a good functionality. Additionally the data browsing and grouping based on meta data was stated to be very informative. More controversial aspects were the actual performance of the AI model, which most expected to be higher, and therefore were surprised. Furthermore the participants were very mixed about the visual explanation method (attribution by occlusion); some said it helped greatly to understand the reasoning of the AI, while other did not find it useful at all. Aspects that were regarded as negative by some was a missing technological explanation on the concepts of image classifiers. Additionally many participants were somewhat confused by the focus on pneumonia, as many stated that X-ray images can give insights on many more pathologies.

	SIPA (pre)	FOST (pre)	SIPA (post)	FOST (post)	ESS	NASA-TLX
N	14	14	14	14	14	14
M	3.88	4.37	4.26	3.80	3.21	6.67
SD	0.77	0.99	0.91	0.81	0.76	2.03
Min	2.00	1.40	1.67	1.80	1.38	4.00
Max	5.17	5.80	5.17	4.80	4.25	11.00

Table 6.2: Descriptive Statistics of all Results

	SIPA (pre)	FOST (pre)	SIPA (post)	FOST (post)	ESS	NASA-TLX
n_g	7	7	7	7	7	7
M	3.67	3.94	3.82	3.44	2.73	6.34
SD	1.01	1.19	1.06	0.94	0.70	1.94
Min	2.00	1.40	1.67	1.80	1.38	4.00
Max	5.17	5.00	4.50	4.20	3.50	10.20

Table 6.3: Descriptive Statistics of guided Results

	SIPA (pre)	FOST (pre)	SIPA (post)	FOST (post)	ESS	NASA-TLX
n_u	7	7	7	7	7	7
M	4.10	4.80	4.71	4.17	3.70	7.00
SD	0.41	0.54	0.45	0.47	0.48	2.21
Min	3.50	4.20	4.00	3.60	3.00	4.00
Max	4.67	5.80	5.17	4.80	4.25	11.00

Table 6.4: Descriptive Statistics of unguided Results

	SS	*df*	*MS*	*F*	*p*	η^2_p
Time	1.016	1	1.016	5.41	.038*	.311
Time ∗ Guidance	0.397	1	0.397	2.11	.172	.150
Residual	2.254	12	0.188			

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 6.5: ANOVA - Within Subjects Effects on SIPA

	SS	*df*	*MS*	*F*	*p*	η^2_p
Guidance	3.11	1	3.11	2.90	.114	.195
Residual	12.86	12	1.07			

Table 6.6: ANOVA - Between Subjects Effects on SIPA

	SS	*df*	*MS*	*F*	*p*	η^2_p
Time	2.2857	1	2.2857	6.4516	.026*	.350
Time ∗ Guidance	0.0229	1	0.229	0.0645	.804	.005
Residual	2.254	12	0.188			

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 6.7: ANOVA - Within Subjects Effects on Trust

	SS	*df*	*MS*	*F*	*p*	η^2_p
Guidance	4.48	1	4.48	4.24	.062	.261
Residual	12.67	12	1.07			

Table 6.8: ANOVA - Between Subjects Effects on Trust

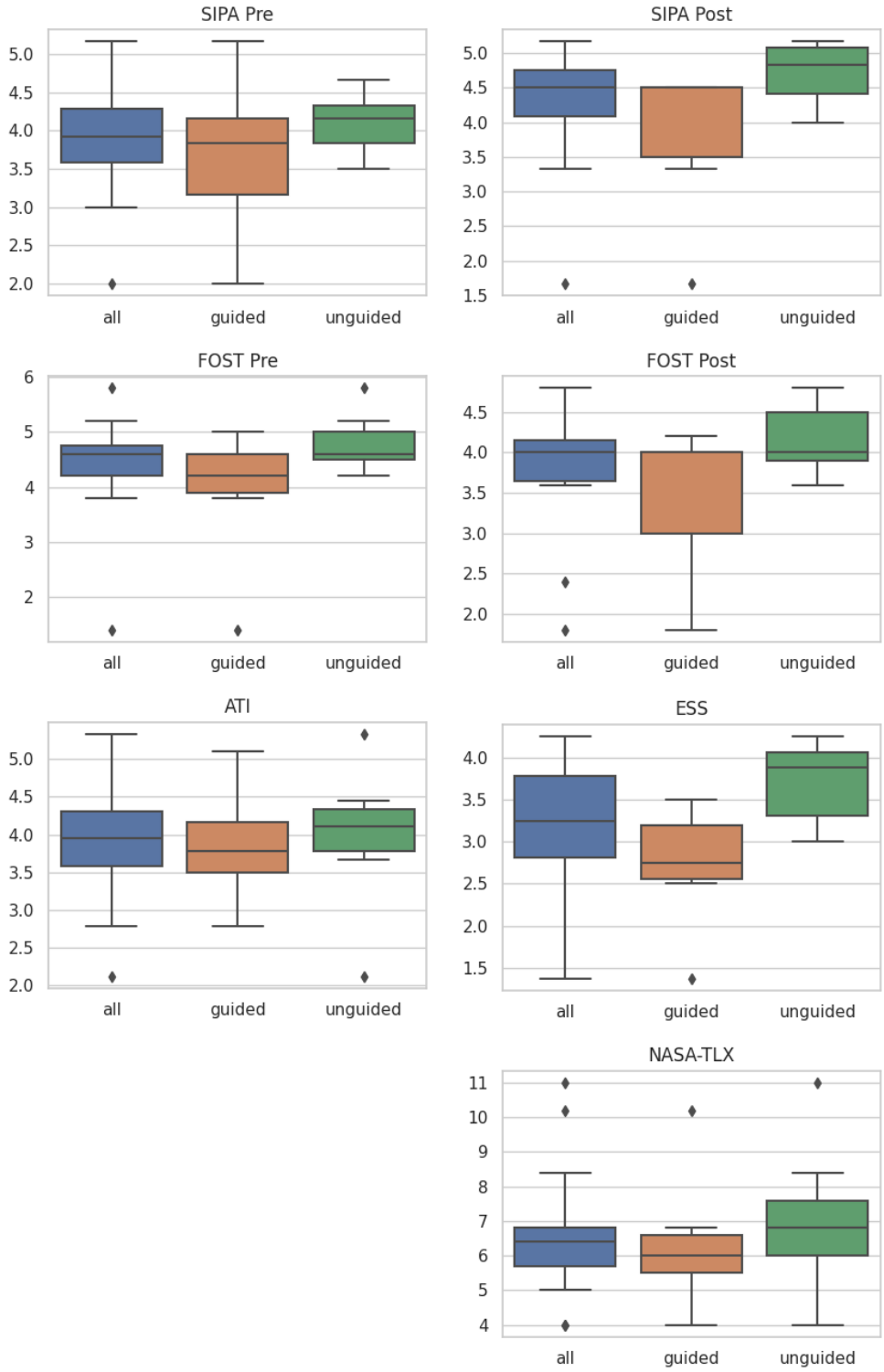


Figure 6.2: Boxplots of Questionnaire Results

		SIPA (pre)	FOST (pre)	ATI
SIPA (pre)	*r*	—		
	p	—		
FOST (pre)	*r*	.57*	—	
	p	.03	—	
ATI	*r*	.13	-.04	—
	p	.66	.89	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 6.9: Correlation Matrix (pre scores)

		ATI	SIPA (post)	FOST (post)	ESS	NASA-TLX
ATI	*r*	—				
	p	—				
SIPA (post)	*r*	.10	—			
	p	.739	—			
FOST (post)	*r*	-.02	.78**	—		
	p	.94	< .01	—		
ESS	*r*	-.03	.71**	.68**	—	
	p	.91	< .01	< .01	—	
NASA-TLX	*r*	.50	-.34	-.38	-.29	—
	p	.07	.23	.18	.31	—

* $p < .05$, ** $p < .01$, *** $p < .001$

Table 6.10: Correlation Matrix (post scores)

6.4 Conclusion on the Evaluation

Through a summative evaluation the effects of the implementation on the user could be measured. An on-site experiment study design including within- and between-subject factors delivered statistically significant results. By surveying $N = 14$ participants with questionnaires such as SIPA, FOST, ESS and NASA-TLX and applying a repeated measure ANOVA, it was discovered how the interaction with an AI assessment system affected explainability and trustworthiness of AI models in the medical context: While SIPA increased, trust decreased. Additionally the correlation between these two variables increased through the interaction. However, the differentiation between guided and unguided interaction did not yield statistically significant results. The subjective workload and the explanation satisfaction was measured to be higher for the unguided group and the explanation satisfaction was correlated with SIPA.

7 Discussion

The following sections give a summarized overview of the thesis, its results implications and limitations, while also discussing directions for further research.

7.1 Summary of Results

The goal of the thesis was to apply human-centered design research on explainability and trustworthiness of black box AI models and therefore studying the effects of HCI to shed light on the requirements for modern AI usage in the medical field.

Through the conception, development and evaluation of an AI assessment system in cooperation with Clearbox, suitable explanation methods were found in order to impact the perceived understandability and trustworthiness of AI models for the specific user group of medical professionals. By studying literature and conducting expert interviews (see chapter 2) requirements, functionalities, interaction- and interface designs for an interactive AI assessment system were conceptualized (see chapter 3). Based on the conception, a functional prototype was implemented with modern web-based technologies in order to accomdate a flexible usage scenario (see chapter 4). By leveraging an open access AI model and data set, the implementation depicts a possible application scenario for the medical domain. Following the human-centered development process, a summative evaluation of the prototype with $N = 14$ participants was conducted. The evaluation consisted of an on-site experiment with medical students including within- and between-subject factors. Analysing the evaluation results yielded insights into the effects of the HCI in both subject groups (see chapter 6).

The statistical analysis of the evaluation results has shown, that the interaction with the assessment system had a significant effect on the subjective understandability (through SIPA) and the perceived trustworthiness (through FOST) of the AI model. However, the

effects go the opposite way: Understandability increased through the interaction, but trustworthiness decreased. This is a interesting finding, because it highlights a particular effect on the user: Through a better subjective understanding, the user's trust in the model was possibly calibrated and the user perceived the model as less trustworthy. Remarkably, the SIPA and FOST pretest scores, which are based only on the textual explanation of the evaluation questionnaire (see Appendix D: Evaluation Questionnaire), were already quite high.

Comparing the between-subject effects, no statistical significance was found. However, the guided group showed a tendency towards less effect size on understandability and trust, compared to the unguided group. This leads to the question, whether an open interaction style leads to a higher perceived feeling of understandability. Sedig et al. (2001) already explored the effects of scaffolding on cognition in learnware, and found that HCI artifacts can extend or, inadvertently, limit human cognition and thought processes. There seems to be a connection between the cognitive load in a learning context and the between-subject effects observed in this evaluation. However, because of the missing significance, the effects of guidance and scaffolding in XAI should be further studied with a focus on the distinction between content-related versus non-content-related aspects of the GUI as proposed by Sedig et al. (2001).

Also the perceived workload was overall higher in the unguided group compared to the guided group. This indicates that the guidance helped with workload, even thought it does not correlate with neither SIPA or trust. However, no meaningful measure of objective workload could be observed. A higher perceived workload in the unguided group, combined with the tendency towards a stronger effect on SIPA and trust, might indicate similar psychological effects as described in the *Elaboration Likelihood Model* by Petty and Cacioppo (1986). If one takes the the rather high ATI score of the participants into consideration, a connection between a higher need for cognition and a stronger effect on SIPA and trust is possible and invites to further research.

While it is not possible to discriminate the effects of each explanation method used by the summative evaluation design, the explanation satisfaction and the SIPA posttest score are also correlated. The display of structured meta data in addition to the image data was part of most explanation methods (and the main focus of one method). The within-subject effect on SIPA leads to the assumption that the presentation of meta data does increase the user's understanding of the AI model's operational range and performance. Referencing the qualitative data, many participants even whished for more meta data,

such as clinical data of the patient, which supports a connection between meta data presentation and explanation satisfaction.

Interestingly ATI, a key facet of user personality, does not have any correlation with the other variables. Seemingly the tendency to actively engage in intensive technology interaction has no effects on the information reception of the user, as neither SIPA or trust are correlated with it. Although the small sample size, only consisting of young medical students with relatively high ATI scores, may distort this finding.

7.2 Implications

The within-subject effects support the answer of the research questions Q1 & Q2: For this AI model and data set, SIPA and trust are correlated. Additionally the correlation increased through the interaction with the assessment system. Therefore the presented functionalities and explanation methods are suitable to increase understanding and possibly optimize the trust levels through the interaction. By supplementing the qualitative data, especially the the input-output experiment and the data browsing functionality were beneficial to the moderate SIPA and trust. On the other hand, the visual explanation was controversially perceived and might not be suitable to explain an AI model to medical professionals. Furthermore the qualitative data shows, that participants from this sample want more technical details about image classifiers.

With the limitation of the missing discrimination of the explanation methods used, research question Q4 can also be answered by the within-subject effects on SIPA: The presentation of meta data in the assessment system is part of the explanation methods used and therefore involved in the significant increase in perceived understandability of the user.

Because of the inconclusiveness of the between-subject ANOVA, research question Q3 cannot be clearly answered: Stronger effects in the unguided group only suggest, that scaffolding and guidance can impede the perceived understandability and therefore the adaption of trust levels. A free, explorative interaction method therefore may be beneficial to understandability and trust.

7.3 Limitations

Originally a more diverse user group was depicted, including medical professionals and data scientists. Already in the user analysis (see section 2.4) first issues arised regarding the quantity of information to be gathered on the user group of data scientists. During the further conception this lack of information led to a strong focus on the user group of medical professionals. The focus was maintained until the summative evaluation, where it culminated in a homogeneous medical participant group. Because of the inherent differences between the user groups, the results are only meaningful for medical professionals. Even then, the participants surveyed belonged to a limited demographic, leading to a more difficult generalizability of the results. Additionally the small sample size might negatively influence the results of the variance analysis.

7.4 Further Research

The inconclusive results on the comparision between guided and unguided interaction only allowed for assumptions regarding the effects on understandabilit and trustworthiness. However this can be a great starting point for further research, focusing on the concrete mode of HCI in the XAI context. By developing more sophisticated user interfaces and interactive explanation methods, the intricacies of guidance and scaffolding for human-AI-interaction could be further explored. Additionally many other feasible explanation methods, such as the omitted comparative explanations or different visual explanations could be implemented and the effects on the users studied. Complemented with a more diverse participant group and additional measures for explainability and trustworthiness, the presented results can be used as a foundation for a further research on HCI in the XAI context.

8 Conclusion

With the all pervasive use of artificial intelligence in current technological advances, the XAI research domain grows ever since. The goal of the thesis was to design and develop an human-centered AI assessment system to make an AI model more explainable and thus facilitate understandability and trustworthiness for the user. Leveraging the knowledge of the HCI domain and combining it with current XAI literature, tools and research partners has resulted in conceptualizing, implementing and evaluating a functional AI assessment system prototype for the context of medical applications. Considering the potential user's needs and requirements throughout the entire development process is an important part of human-centered application development. Expert interviews gave insights into the needs and requirements of the potential users of an AI model in the medical context. Based on these requirements and XAI literature, functionalitites, explanation methods, interaction- and interface design were conceptualized and implemented with a strong focus on understandability and coactive HCI. The functional prototype was then evaluated in a study with medical students. The evaluation yielded results, which depict clear connections between interactive explanations, understandability and trustworthiness of AI models. A summative evaluation of a AI assessment system prototype has shown that user's subjective understanding of the AI model increased through the interaction with said system. Furthermore the user's trust has decreased through the interaction with the system for the specific AI model used. Therefore interactive explanation methods, such as (1) contextual train data browsing (2) showing model limitations through false positives and false negatives (3) input-output experiments (4) visual explanation by attribution through occlusion are suitable for moderating the user's subjective understanding and perceived trustworthiness of an AI model. In addition it was found that guidance in the HCI reduced the explanation satisfaction, while showing no significant effects on understandability and trust for the users sampled. However, the results give clear insights into the complex concepts of understandability and trustworthiness for human-AI-interaction in the medical context and create a foundation for further research.

List of Figures

List of Tables

List of Source Codes

Sources

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *arXiv:1910.10045 [cs]*. Retrieved May 19, 2021, from http://arxiv.org/abs/1910.10045

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 258–262. https://doi.org/10.1145/3301275.3302289

Clearbox AI. (2021a). *Clearbox AI Model Assessment* (Whitepaper). Retrieved April 2021, from https://clearbox.ai/pdf/ClearboxAI_Technical_Whitepaper.pdf

Cohen, J. P., Hashir, M., Brooks, R., & Bertrand, H. (2020). On the limits of cross-domain generalization in automated x-ray prediction. *arXiv:2002.02497 [cs, eess, q-bio, stat]*. Retrieved October 25, 2021, from http://arxiv.org/abs/2002.02497

Cohen, J. P., Viviano, J. D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M. P., Chaudhari, A., Brooks, R., Hashir, M., & Bertrand, H. (2021). TorchXRayVision: A library of chest x-ray datasets and models. *arXiv:2111.00595 [cs, eess]*. Retrieved December 9, 2021, from http://arxiv.org/abs/2111.00595

DIN EN ISO 9241-210: Ergonomie der Mensch-System-Interaktion - Teil 210: Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme. (2011). DIN Deutsches Institut für Normung e. V. Beuth Verlag.

European Commision. (2020). *On Artificial Intelligence - A European approach to excellence and trust* (Whitepaper). Retrieved September 2021, from https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction*, *35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150

Franke, T., Trantow, M., Günther, M., Krems, J. F., Zott, V., & Keinath, A. (2015). Advancing electric vehicle range displays for enhanced user experience: The relevance of trust and adaptability. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 249–256. https://doi.org/10.1145/2799250.2799283

Gediga, G., Hamborg, K.-C., & Düntsch, I. (2002). Evaluation of software systems. *Encyclopedia of computer science and technology*, *45*(supplement 30), 127–53.

Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T., & Bernstein, M. S. (2021). The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445423

Hart, S. G. (2006). Nasa-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Harte, R., Glynn, L., Rodríguez-Molinero, A., Baker, P. M., Scharf, T., Quinlan, L. R., & ÓLaighin, G. (2017). A human-centered design methodology to enhance the usability, human factors, and user experience of connected health systems: A three-phase methodology. *JMIR Human Factors*, *4*(1), e8. https://doi.org/10.2196/humanfactors.5443

Hering, E. (1984). Programmablaufplan nach DIN 66001. In E. Hering (Ed.), *Software-Engineering: Mit 77 Bildern und 22 Übungsaufgaben* (pp. 26–34). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-86222-8_4

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2019). Metrics for explainable AI: Challenges and prospects. *arXiv:1812.04608 [cs]*. Retrieved May 19, 2021, from http://arxiv.org/abs/1812.04608

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence

in joint activity. *Journal of Human-Robot Interaction*, *3*(1), 43–69. https://doi.org/10.5898/JHRI.3.1.Johnson

Keane, M. T., & Kenny, E. M. (2019). How case based reasoning explained neural networks: An XAI survey of post-hoc explanation-by-example in ANN-CBR twins. *arXiv:1905.07186 [cs]*, *11680*, 155–171. https://doi.org/10.1007/978-3-030-29249-2_11

Knapič, S., Malhi, A., Saluja, R., & Främling, K. (2021). Explainable artificial intelligence for human decision-support system in medical domain. *arXiv:2105.02357 [cs]*. Retrieved June 11, 2021, from http://arxiv.org/abs/2105.02357

Kopetz, H. (1976). Functional specification. In H. Kopetz (Ed.), *Software reliability* (pp. 33–38). Macmillan Education UK. https://doi.org/10.1007/978-1-349-86129-3_5

Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., & Samek, W. (2016). The lrp toolbox for artificial neural networks. *Journal of Machine Learning Research*, *17*(114), 1–5. http://jmlr.org/papers/v17/15-618.html

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1096. https://doi.org/10.1038/s41467-019-08987-4

Meske, C., & Bunde, E. (2020). Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. *arXiv:2002.01543 [cs]*, *12217*, 54–69. https://doi.org/10.1007/978-3-030-50334-5_4

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv:1902.01876 [cs]*. Retrieved May 19, 2021, from http://arxiv.org/abs/1902.01876

Nielsen, H., Mogul, J., Masinter, L. M., Fielding, R. T., Gettys, J., Leach, P. J., & Berners-Lee, T. (1999). Hypertext Transfer Protocol – HTTP/1.1. https://doi.org/10.17487/RFC2616

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Communication and persuasion* (pp. 1–24). Springer.

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. *arXiv:1803.07517 [cs, stat]*. Retrieved May 19, 2021, from http://arxiv.org/abs/1803.07517

Ras, G., Xie, N., van Gerven, M., & Doran, D. (2021). Explainable deep learning: A field guide for the uninitiated. *arXiv:2004.14545 [cs, stat]*. Retrieved October 6, 2021, from http://arxiv.org/abs/2004.14545

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *arXiv:1602.04938 [cs, stat]*. Retrieved May 19, 2021, from http://arxiv.org/abs/1602.04938

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). Retrieved June 24, 2021, from https://ojs.aaai.org/index.php/AAAI/article/view/11491

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

Schrills, T., Zoubir, M., Bickel, M., Kargl, S., & Franke, T. (2021). Are Users in the Loop? Development of the Subjective Information Processing Awareness Scale to Assess XAI.

Sedig, K., Klawe, M., & Westrom, M. (2001). Role of interface manipulation style and scaffolding on cognition and concept learning in learnware. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *8*(1), 34–59.

Soloway, E., Guzdial, M., & Hay, K. E. (1994). Learner-centered design: The challenge for HCI in the 21st century. *interactions*, *1*(2), 36–48.

Trigg, R. H. (1988). Guided tours and tabletops: Tools for communicating in a hypertext environment. *ACM Transactions on Information Systems (TOIS)*, *6*(4), 398–414.

Vogt, W. P., & Johnson, B. (2011). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences*. Sage.

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2016). *Engineering Psychology and Human Performance*. Routledge.

Zeiler, M. D., & Fergus, R. (2013). Visualizing and understanding convolutional networks. *arXiv:1311.2901 [cs]*. Retrieved June 10, 2021, from http://arxiv.org/abs/1311.2901

Websites

Clearbox AI. (2021b, October). *Manage AI models with confidence*. https://clearbox.ai/

CoCoAI. (2021, September). *Cooperative and communicating AI methods for medical image-guided diagnostics - A research project at the University of Lübeck.* https://cocoai.uni-luebeck.de

Cohen, J. P., Viviano, J., Morrison, P., Brooks, R., Hashir, M., & Bertrand, H. (2021, October). *TorchXRayVision: A library of chest X-ray datasets and models.* https://github.com/mlmed/torchxrayvision

European Commision. (2021, March). *Ethics guidelines for trustworthy AI.* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Facebook. (2021, July). *Captum: Model Interpretability for PyTorch.* https://captum.ai/

Google. (2021, June). *People + AI Guidebook.* https://pair.withgoogle.com/guidebook/

J. J. Garrett. (2021, August). *The Elements of User Experience.* http://www.jjg.net/elements/pdf/elements.pdf

NumPy. (2021, October). *NumPy is the fundamental package needed for scientific computing with Python.* https://github.com/numpy/numpy

Pandas. (2021, October). *Powerful Python Data Analysis Toolkit.* https://github.com/pandas-dev/pandas

PyTorch. (2021, October). *An open source machine learning framework that accelerates the path from research prototyping to production deployment.* https://pytorch.org/

Radiological Society of North America. (2021, July). *RSNA Pneumonia Detection Challenge.* https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

Streamlit Inc. (2021a, July). *Streamlit: The fastest way to build and share data apps.* https://streamlit.io/

ThinkSono. (2021, September). *The world's first software to detect DVT!* https://thinksono.com/

TIOBE Software BV. (2021, October). *TIOBE Index for November 2021.* https://www.tiobe.com/tiobe-index/python/

Software

DeepL GmbH. (2021, June). *DeepL Translator.* https://www.deepl.com/translator

Streamlit Inc. (2021b, October). *Streamlit: The fastest way to build and share data apps* (Version 1.2.0). https://github.com/streamlit/streamlit

Abbreviations

AI Artificial Intelligence

AAII Affinity for AI Interaction

ANOVA Analysis of Variance

ATI Affinity for Technology Interaction

AUROC Area Under the Receiver Operating Characteristics

API Application Programming Interface

CNN Convolutional Neural Network

DNN Deep Neural Network

DVT Deep Vein Thrombosis

ESS Explanation Satisfaction Scale

FOST Facets of System Trustworthiness

GUI Graphical User Interface

HCI Human-Computer-Interaction

ML Machine Learning

NASA-TLX NASA Task Load Index

SIPA Subjective Information Processing Awareness

XAI Explainable Artificial Intelligence

Glossary

Analysis of Variance Collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among means.

Artificial Intelligence Systems that display intelligent behaviour by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals.

Affinity for Technology Interaction Tendency to actively engage in intensive technology interaction, as a key personal resource for coping with intelligent systems.

Area Under the Receiver Operating Characteristics Important evaluation metric for checking any classification model's performance.

Black Box System which can be viewed in terms of its inputs and outputs (or transfer characteristics), without any knowledge of its internal workings.

Convolutional Neural Network Convolutional neural networks are a class of artificial neural networks, most commonly used for processing structured arrays of data such as images..

Deep Neural Network Deep neural networks are a powerful category of machine learning algorithms implemented by stacking layers of neural networks along the depth and width of smaller architectures.

Deep Vein Thrombosis Deep vein thrombosis occurs when a blood clot forms in one or more of the deep veins in your body, usually in the legs.

Explainable Artificial Intelligence Artificial intelligence in which the results of the solution can be understood by humans.

Machine Learning Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Appendices

Appendix A: DVD Contents

Oft ein Default: Was findet man auf der beiliegenden DVD in welchem Verzeichnis? Max. 1 Seite.

In jedem Fall die PDF der Arbeit, den Programmcode, Daten (anonymisiert!).

Niemals Interviewaufzeichnungen, Einverständniserklärungen oder ähnliche personenbezogene Daten auf die DVD brennen — Sie haben in den meisten Fällen Anonymität zugesichert und die DVD ist frei zugänglich (ein Exemplar der Arbeit kommt in die Bibliothek).

Text ...

Appendix B: Interview Guideline

Datum:_____
ID: _____

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE UND INTERAKTIVE SYSTEME

Interviewleitfaden zur Nutzung, Vertrauenswürdigkeit und Erklärbarkeit von künstlicher Intelligenz

Einführung

Begrüßung:

Hallo, vielen Dank, dass Sie sich Zeit genommen haben für dieses Interview. Ich bin Philipp Bzdok und studieren Medieninformatik im Master am Institut für Multimediale und Interaktive Systeme der Universität zu Lübeck. Unterstützt werde ich heute von Mona Bickel.

Erläuterung:

Wie Sie vielleicht schon wissen, möchte ich im Rahmen meiner Abschlussarbeit die Anforderungen und Bedürfnisse von Menschen für den professionellen Umgang mit KI-Systemen beleuchten. KI-Systeme sind in diesem Kontext Computer Systeme, welche durch den Einsatz von maschinellem Lernen komplexe Aufgaben lösen können (z.B. Erkennung von TVT in Ultraschall Bildern). Dies ist der erste wichtige Schritt für die Weiterentwicklung bestehender KI-Assessment-Systeme zur Förderung von Erklärbarkeit und Vertrauenswürdigkeit von KI in medizinischen Anwendungen. Denn nur wenn ein KI-System erklärbar und vertrauenswürdig ist, können auf Basis der Vorhersagen fundierte Entscheidungen im Arbeitsalltag getroffen werden.

Plan:

Heute möchte ich in diesem Interview herausfinden, was Ihre Erfahrungen, Bedürfnisse und Anforderungen bzgl. KI-Systemen im wissenschaftlichen / medizinischen (Arbeits-) Kontext sind. Hierbei geht es in erster Linie um Ihre subjektive Meinung und Erfahrung! Abschließend möchten ich noch ihre allgemeine Affinität zu KI Interaktion durch einen kurzen Fragebogen ermitteln.

Datenschutz und Einwilligung:

Wir würden das Interview für eine bessere Möglichkeit zur Auswertung gerne aufzeichnen. Die aufgezeichneten Daten werden anonymisiert für die Abschlussarbeit ausgewertet.
Ist das für Sie in Ordnung?

Start der Aufzeichnung
Ziel: 2 Minuten pro Frage + Puffer

Datum:_____
ID: _____

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE UND INTERAKTIVE SYSTEME

Mediziner

1 Fragen zur Person

1.1 Demografie

1.1.1 Alter _____

1.1.2 Geschlecht _____

1.1.3 Beruf _____

1.1.4 Bildungsstand _____

1.2 Welche KI Systeme in Ihrem fachlichen Umfeld kennen Sie? (Frage zum Warmwerden)

_____

_____

_____

_____

1.3 Welche KI-Systeme benutzen Sie im Arbeitsalltag? (Gebrauch von KI)

_____

_____

_____

_____

1.3.1 Nachfrage, wenn Nutzung gegeben: Wie **zufrieden** waren Sie mit dem Gebrauch des KI-Systems? (Nachfrage zum Gebrauch)

_____

_____

_____

_____

Datum:_____
ID: _____

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE UND INTERAKTIVE SYSTEME

1.4 Welche **Vorteile** sehen Sie in Ihrem Arbeitsalltag beim Gebrauch von KI? (Perspektive zur KI Nutzung)

_____

_____

_____

_____

1.5 Welche **Risiken** sehen Sie in Ihrem Arbeitsalltag beim Gebrauch von KI? (Perspektive zur KI Nutzung)

_____

_____

_____

_____

Bitte denken Sie an einen konkreten Anwendungsfall von KI aus ihrem Arbeitsalltag und beschreiben Sie diesen kurz! (Alternativ: AutoDVT Use-Case etablieren) (Use-Case für weitere Fragen etablieren)

_____

_____

_____

_____

1.6 Inwiefern **vertrauen** Sie den Ergebnissen des KI-Systems (aus dem Use-Case)? (Vertrauen zu KI)

_____

_____

_____

_____

Datum:_____
ID: _____

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE UND INTERAKTIVE SYSTEME

1.6.1 Nachfrage: Aufgrund welcher **Informationen** bzw. Interaktion machen Sie dieses Vertrauen fest?

_____

_____

_____

_____

1.6.2 Nachfrage: Was waren Schlüsselmomente, die Ihr Vertrauen **verändert** haben?

_____

_____

_____

_____

1.7 Welche **Fragen** stellen Sie sich, wenn sie das System benutzen? (Potenzielle Probleme bei der KI Nutzung)

_____

_____

_____

_____

1.8 Wie gehen Sie vor, um die Funktionsweise des Systems zu **verstehen**? (Eigene Erklärungsmethoden)

_____

_____

_____

_____

Datum:_____
ID: _____

1.9 Was verstehen Sie unter dem Begriff „Erklärbare Künstliche Intelligenz“ oder auch „Explainable Artificial Intelligence“ (XAI)? (Vertrautheit mit XAI)

1.9.1 Nachfrage: Inwiefern könnte eine KI-Modell-Erklärung Ihr Vertrauen in die Vorhersagen des Systems verändern?

2 Fragen zur Interaktion mit KI-Assessment-Systemen

Es ist ein System in der Entwicklung welches eine umfangreiche KI-Modell Bewertung erstellt. Eine solche Bewertung enthält Informationen über Veränderungen der Daten über die Zeit und besondere Randfälle. Des Weiteren zeigt es Modellgrenzen auf und bewertet die Robustheit und Interpretierbarkeit des Modells. Dazu werden diverse Erklärungsmethoden und Visualisierungen genutzt. (Einleitung des AI-Assessment Systems [Clearbox])

2.1 Inwiefern könnten konkrete Beispiele (auch Local Explanations genannt) Ihnen helfen die Ergebnisse des KI-Modells besser zu **verstehen**? (Einschätzung zu Local Explanations)

Datum:_____
ID: _____

2.2 Wann würden konkrete Beispiele Ihnen **nicht** helfen die Ergebnisse des KI-Modells besser zu **verstehen**? (Einschätzung zu Local Explanations)

_____

_____

_____

_____

2.3 Wenn Sie das KI-System etwas Fragen könnten, um es besser **verstehen** zu können, was wäre es? (Primärfrage, Zeit lassen! Informationsverarbeitung)

_____

_____

_____

_____

2.4 Was würden Sie sich für die KI-Forschung **wünschen**, damit solche Systeme in Ihrem medizinischen Kontext mehr Anwendung finden können? (Vertrauen – Verhalten Verbindung)

_____

_____

_____

_____

2.5 Angenommen ein KI-System ist 100-prozentig zuverlässig, bräuchten Sie dann noch Erklärungen? (Zuverlässigkeit vs. Vertrauen vs. Verständnis)

_____

_____

_____

_____

Datum:_____
ID: _____

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE UND INTERAKTIVE SYSTEME

3 Abschluss

3.1 Haben Sie über die unsere Gesprächsinhalte noch weitere Anmerkungen oder Fragen?

_____

_____

_____

_____

4 AKII (ATI) Fragebogen per Limesurvey ausfüllen lassen.
(https://bzdok.limesurvey.net/924128?lang=en)

Datum:_____
ID: _____

UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE UND INTERAKTIVE SYSTEME

AI Researchers / Data Scientists

1 Fragen zur Person

1 Demografie

1.1.1 Alter _____

1.1.2 Geschlecht _____

1.1.3 Beruf _____

1.1.4 Bildungsstand _____

1.2 Welche Machine-Learning-Modelle nutzen Sie in Ihrem Arbeitsalltag? (Frage zum Warmwerden)

_____

_____

_____

_____

1.3 Wie wählen Sie zwischen verschiedenen Machine-Learning-Modellen? (Gebrauch von KI)

_____

_____

_____

_____

1.4 Was sind die wichtigsten Aspekte bei der Auswahl eines konkreten Modells? (Vergleich von KI-Modellen)

_____

_____

_____

_____

Datum:_____
ID: _____

1.5 Welche **Vorteile** sehen Sie beim Gebrauch von KI? (Perspektive zur KI-Nutzung)

_____

_____

_____

_____

1.6 Welche **Risiken** sehen Sie beim Gebrauch von KI? (Perspektive zur KI-Nutzung)

_____

_____

_____

_____

Bitte denken Sie an einen konkreten Anwendungsfall von KI aus ihrem Arbeitsalltag und beschreiben Sie diesen kurz! (Alternativ: AutoDVT Use-Case etablieren) (Use-Case für weitere Fragen etablieren)

_____

_____

_____

_____

1.7 Inwiefern **vertrauen** Sie den Ergebnissen des KI-Systems (aus dem Use-Case)? (Vertrauen zu KI)

_____

_____

_____

_____

Datum:_____
ID: _____

1.7.1 Nachfrage: Aufgrund welcher **Informationen** bzw. Interaktion machen Sie dieses Vertrauen fest?

_____

_____

_____

_____

1.7.2 Nachfrage: Was waren Schlüsselmomente, die Ihr Vertrauen **verändert** haben?

_____

_____

_____

_____

1.8 Welche **Fragen** stellen Sie sich, wenn sie das System benutzen? (Potenzielle Probleme bei der KI Nutzung)

_____

_____

_____

_____

1.9 Wie gehen Sie vor, um die Funktionsweise des Systems zu **verstehen**? (Eigene Erklärungsmethoden)

_____

_____

_____

_____

Datum:_____
ID: _____

1.10 Welche (XAI) Methoden zur Erklärbarkeit von KI haben Sie **bereits genutzt**? (Vertrautheit mit XAI)

1.10.1 Nachfrage für jede Methode: Inwiefern hat die XAI Methode ihr Vertrauen in die Vorhersagen des Modells **beeinflusst**?

2 Fragen zur Interaktion mit KI-Assessment-Systemen

Es ist ein System in der Entwicklung welches eine umfangreiche KI-Modell Bewertung erstellt. Eine solche Bewertung enthält Informationen über Veränderungen der Daten über die Zeit und besondere Randfälle. Des Weiteren zeigt es Modellgrenzen auf und bewertet die Robustheit und Interpretierbarkeit des Modells. Dazu werden diverse Erklärungsmethoden und Visualisierungen genutzt. (Einleitung des AI-Assessment Systems [Clearbox])

2.1 Wann brauchen Sie beispielhafte, **lokale Erklärungen** eines Machine-Learning-Modells? (Local Explanations Bedarf)

Datum:_____
ID: _____

2.2 Wann brauchen Sie **global valide Erklärungen** eines Machine-Learning-Modells? (Global Explanation Bedarf)

2.3 Wenn Sie das KI-System etwas Fragen könnten, um es besser **verstehen** zu können, was wäre es? (Primärfrage, Zeit lassen! Informationsverarbeitung)

2.4 Was würden Sie sich für die KI-Forschung **wünschen**, damit solche Systeme in Ihrem fachlichen Kontext mehr Anwendung finden können? (Vertrauen – Verhalten Verbindung)

2.5 Angenommen ein KI-System ist 100-prozentig zuverlässig, bräuchten Sie dann noch Erklärungen? (Zuverlässigkeit vs. Vertrauen vs. Verständnis)

Datum:_____
ID: _____

3 Abschluss

3.1 Haben Sie über die unsere Gesprächsinhalte noch weitere Anmerkungen oder Fragen?

_____

_____

_____

_____

4 AKII (ATI) Fragebogen per Limesurvey ausfüllen lassen.
(https://bzdok.limesurvey.net/924128?lang=en)

Appendix C: Interaction Flowcharts

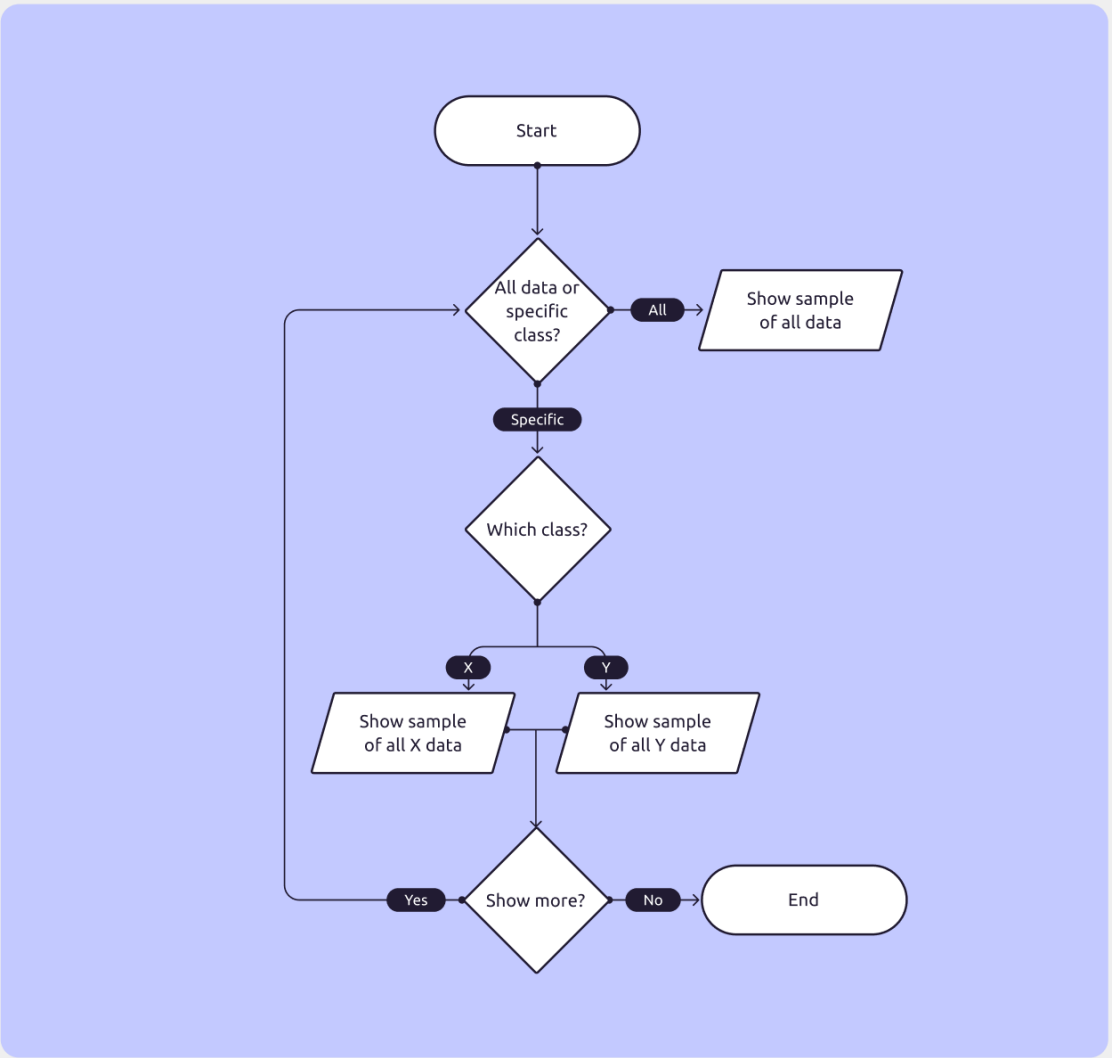


Figure 8.1: Flowchart - Browse Training Data for a given Class

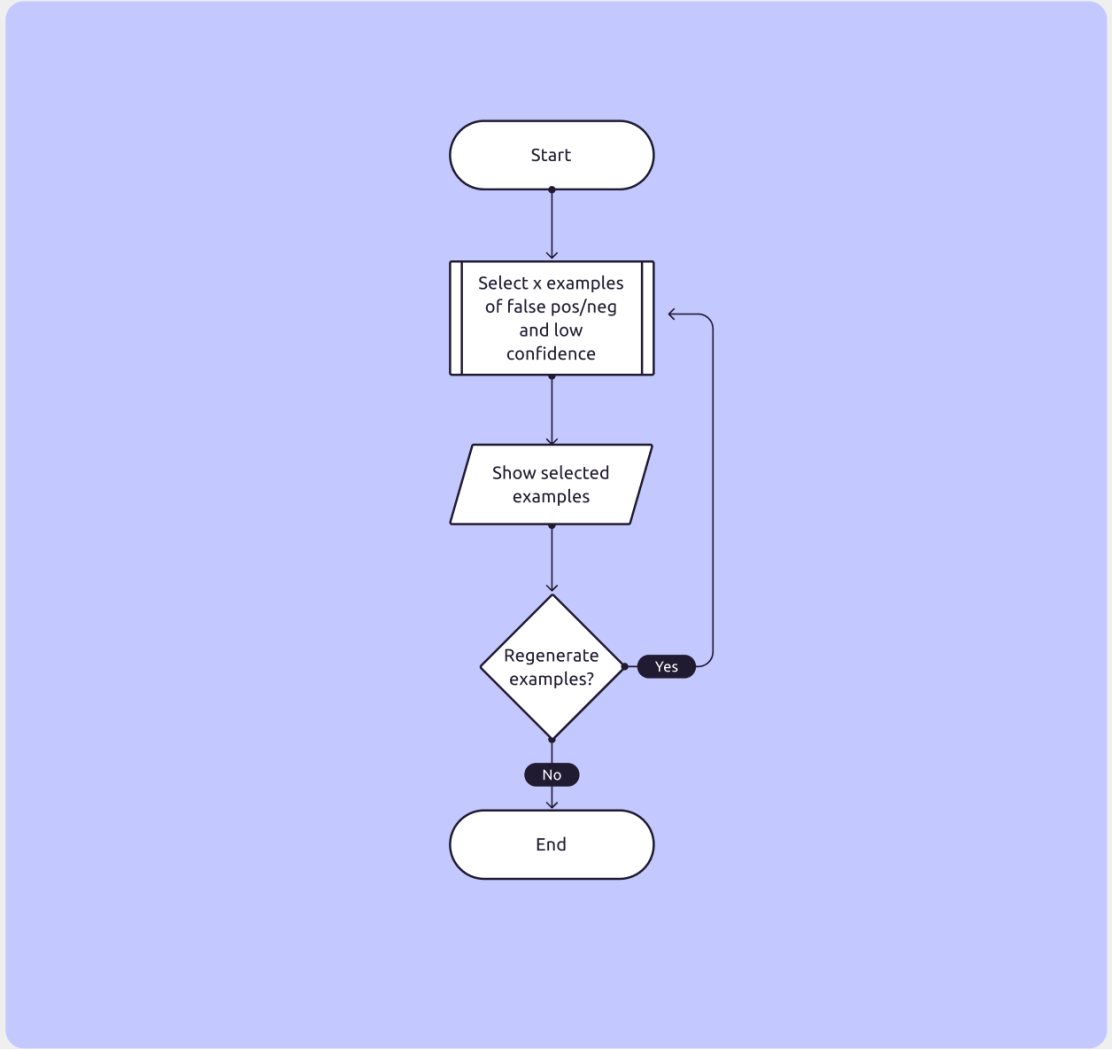


Figure 8.2: Flowchart - Show Examples of false positive / negative or low confidence

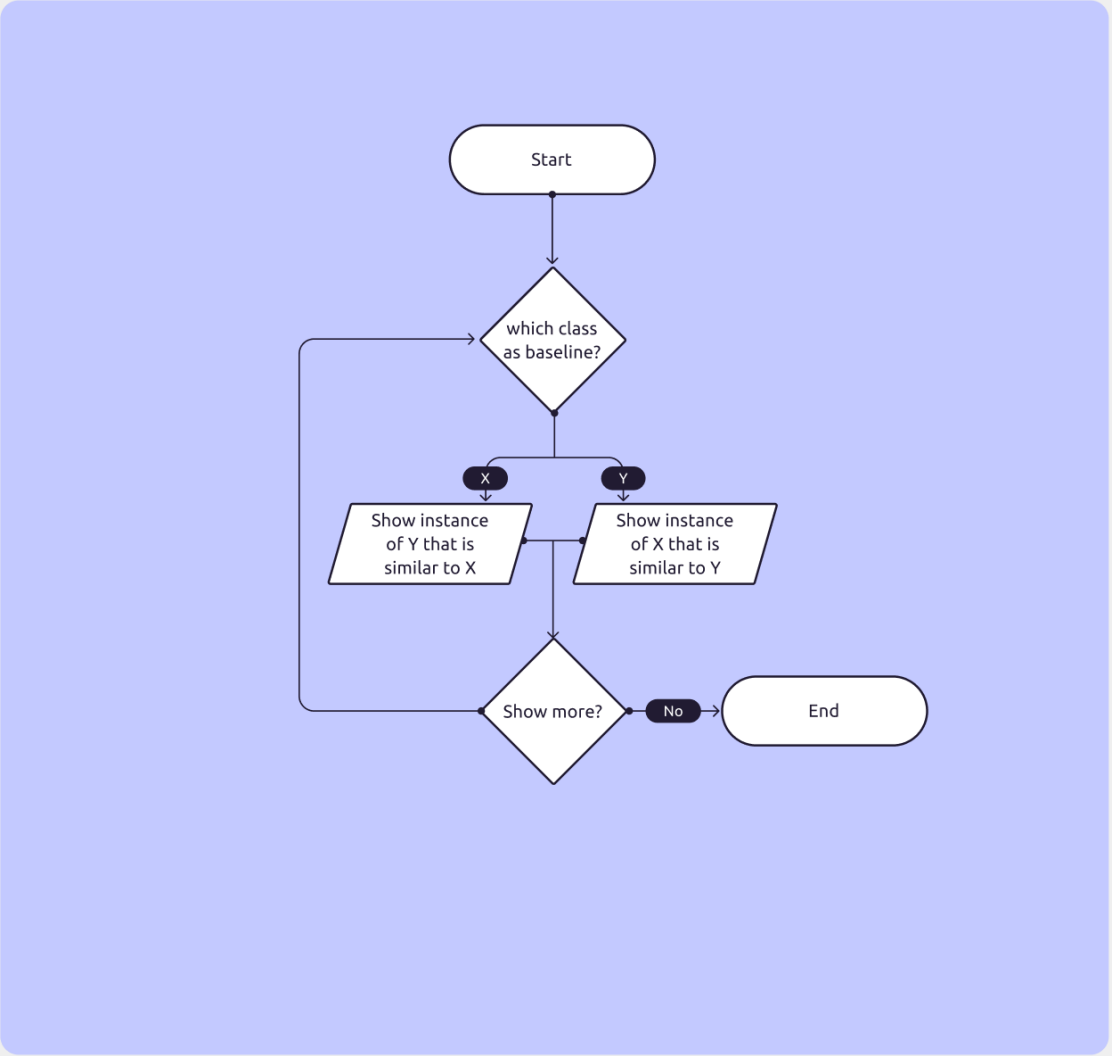


Figure 8.3: Flowchart - Show Similarities

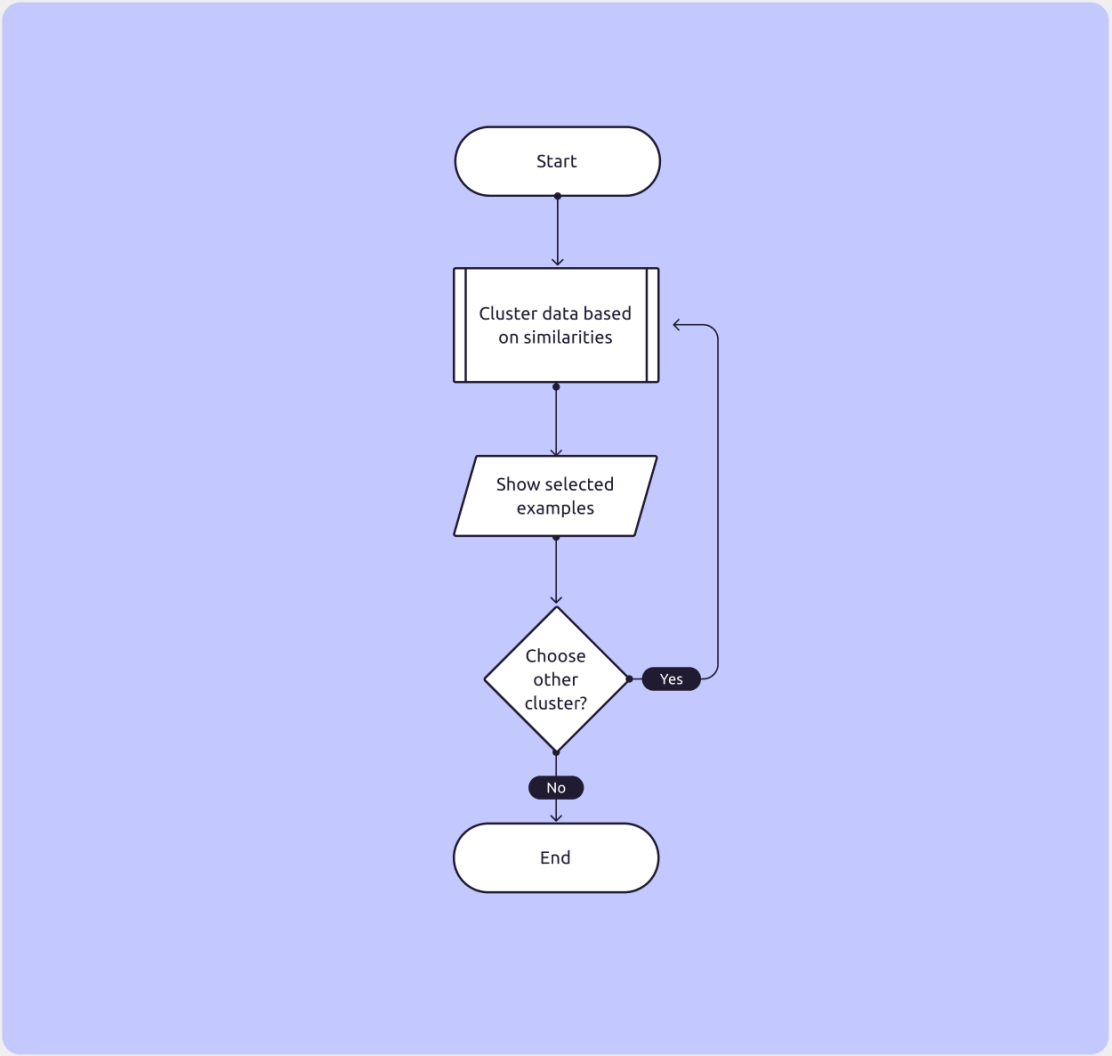


Figure 8.4: Flowchart - Grouping of Data based on Similarities

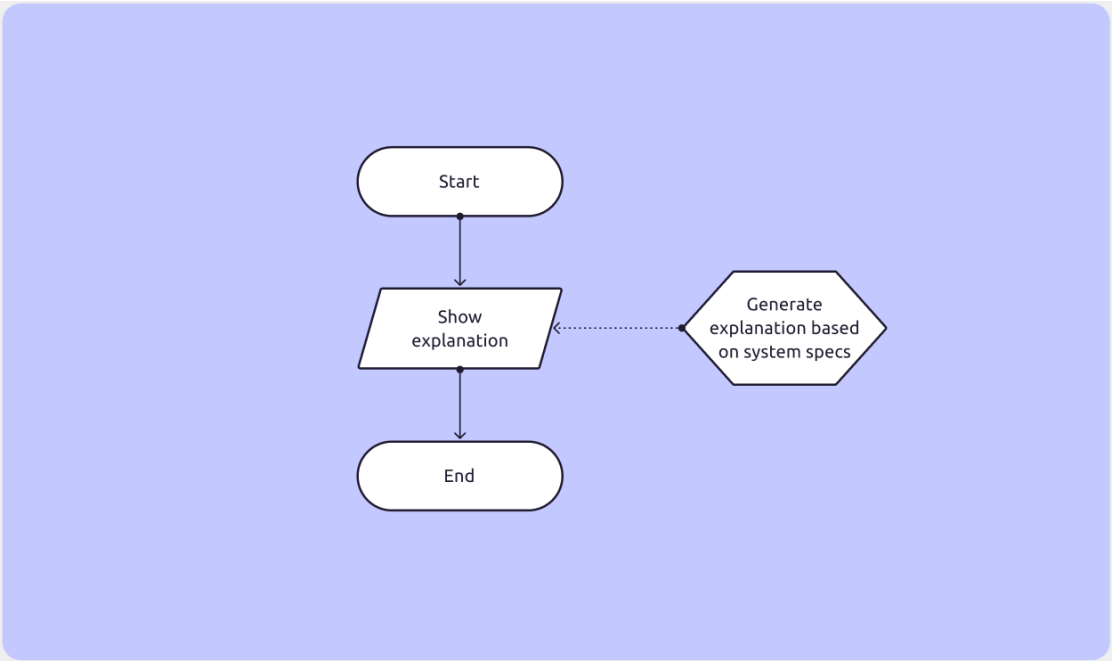


Figure 8.5: Flowchart - Overview of General System Capabilities

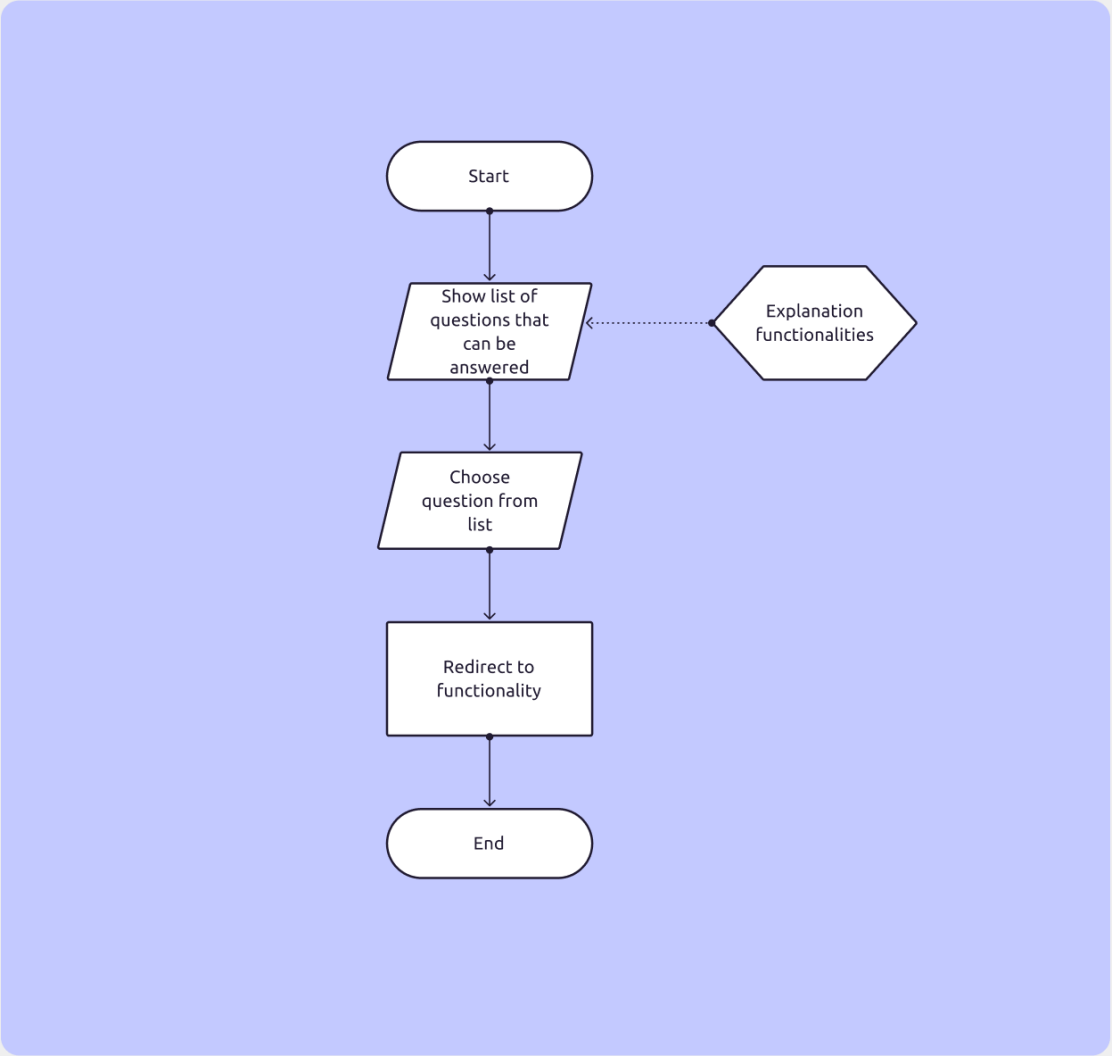


Figure 8.6: Flowchart - Show Written Explanations via Templates

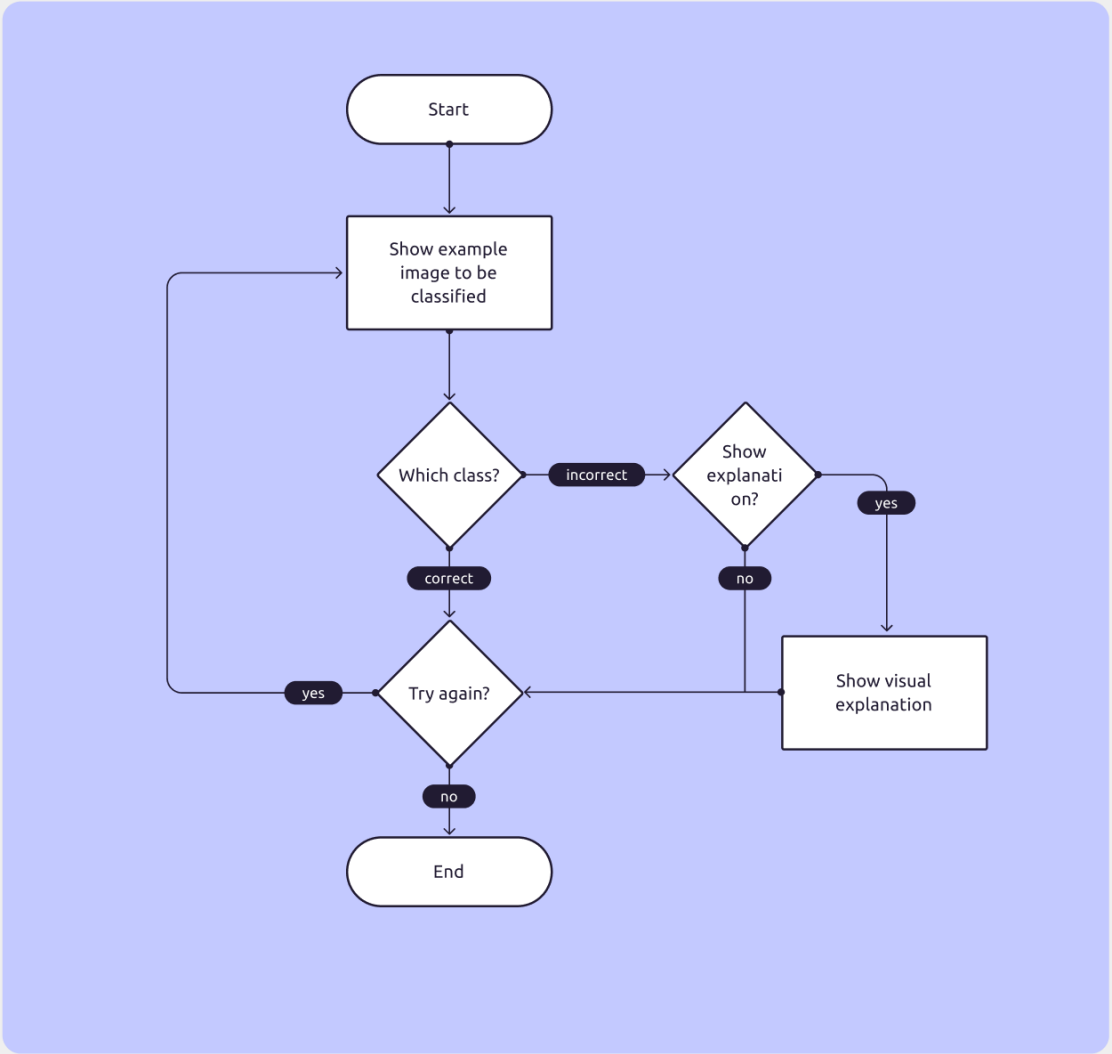


Figure 8.7: Flowchart - Input-Output Experiment

Appendix D: Evaluation Questionnaire

ID: _____ Datum: _____

Evaluation eines KI Assessment Systems im Rahmen einer Masterarbeit

Einleitung

Vielen Dank, dass Sie sich für die Evaluation Zeit genommen haben!
Wie Sie vielleicht schon wissen, geht es heute um künstliche Intelligenz (**KI**). Auch im medizinischen Bereich ist KI stark im Trend, insbesondere bei bildgebenden Verfahren: KI kann zum Beispiel medizinischen Fachpersonal dabei helfen Thrombose in Ultraschallbildern zu erkennen. Damit solche KI-Systeme jedoch im klinischen Alltag verwendet werden können, müssen diese erklärbar sein damit auch die Nutzenden ihnen vertrauen und basierend auf den Ergebnissen der KI entscheidungen treffen können. Würden Sie einem solchen System vertrauen, oder gar für sich entscheiden lassen?

Im Rahmen einer Masterarbeit wurde ein System entwickelt, um KI erklärbarer und vertrauenswürdiger zu machen. Dieses System gilt es zu evaluieren.

Es folgt eine allgemeine Erklärung zur KI, den gängigen Metriken & dem Assessment (deut. *Einschätzung*) System. Daraufhin soll ein erster Fragebogen ausgefüllt werden. Nachdem Sie den Fragebogen ausgefüllt haben startet der interaktive Teil der Evaluation. Dabei geht es für Sie darum das System auf Ihre persönliche Art zu bedienen. Ziel der Interaktion ist es, die KI besser zu verstehen. Sie haben 15 Minuten Zeit das System zu erkunden und so über die Fähigkeiten, Stärken & Schwächen der KI zu lernen. Wenn Sie denken, keine weiteren Erkenntnisse aus der Interaktion mehr zu gewinnen, können Sie die Interaktion auch früher stoppen. Abschließend soll noch ein Fragebogen ausgefüllt werden.

Falls Sie während der Durchführung Fragen haben, dürfen Sie den Versuchsleiter jederzeit um Unterstützung bitten.

ID: _____ Datum: _____

Erklärungen

Künstliche Intelligenz

Ein KI, so wie Sie hier verwendet wird, kann anhand von Bilddaten eine Klassifizierung vorzunehmen. Konkret ist das hier verwendete KI-Modell in der Lage anhand von digitalen Röntgenbildern eine Einschätzung zu treffen, ob eine Person unter Lungenentzündung leidet. Die KI wurde mit zehntausenden Röntgenbildern, welche aus unterschiedlichsten Quellen stammen, trainiert und kann anhand eines Bildes und innerhalb einer Sekunde entscheiden, ob es sich um einen pathologischen Befund handelt. Die hierfür verwendete Technologie heißt “Densely Connected Convolutional Network”; Wissenschaftler haben herausgefunden, dass diese Art von KI besonders geeignet ist um Bilder zu analysieren und zu klassifizieren.

Metriken

Wie auch andere Systeme und Prozesse können KI-Modelle durch Metriken beschrieben werden. Die hier verwendeten Metriken sind:

Accuracy (deut. *Vertrauenswahrscheinlichkeit*): Gibt den Anteil aller Objekte an, die korrekt klassifiziert werden.

Precision (deut. *Genauigkeit*): Gibt den Anteil der korrekt als positiv klassifizierten Ergebnisse an der Gesamtheit der als positiv klassifizierten Ergebnisse an. Beispielsweise gibt die Genauigkeit eines medizinischen Tests an, welcher Anteil der Personen mit positivem Testergebnis auch tatsächlich krank ist.

Recall (deut. *Sensitivität*): Gibt die Wahrscheinlichkeit an, mit der ein positives Objekt korrekt als positiv klassifiziert wird. Beispielsweise entspricht die Sensitivität bei einer medizinischen Diagnose dem Anteil an tatsächlich Kranken, bei denen die Krankheit auch erkannt wurde.

F1: Das F-Maß kombiniert Genauigkeit und Sensitivität mittels des gewichteten harmonischen Mittels.

Assessment System

Das (Assessment)-System ermöglicht einen Einblick in verschiedene Aspekte der KI. Es werden allgemeine Informationen bereitgestellt und Standard-Metriken verwendet um das KI-Modell allgemein zu beschreiben. Darüber hinaus ist es möglich die Trainingsdaten zu sichten und spezielle Randfälle innerhalb der Daten zu erkunden. Des weiteren kann man durch ein Eingabe-Ausgabe-Experiment seine eigene Einschätzung über ein Röntgenbild mit der Einschätzung der KI vergleichen und zusätzlich visualisieren lassen, anhand welcher Bildbereiche die KI zu einer Entscheidung gekommen ist.

ID: _____ Datum: _____

Datenschutz

Bevor wir beginnen, benötigen wir Ihr explizites Einverständnis, dass wir Ihre Daten speichern, auswerten und verarbeiten dürfen. Es folgt die Datenschutz- und Einwilligungserklärung.

Bei Fragen kontaktiere Sie bitte den Studienleiter Philipp Bzdok (bzdok@imis.uni-luebeck.de).

Datenschutz- und Einwilligungserklärung
Die Teilnahme an der Studie sowie die Einwilligung in die Verarbeitung von personenbezogenen Daten ist freiwillig. Auch nach erteilter Einwilligung kann die Teilnahme jederzeit und ohne Angabe von Gründen beendet werden, ohne dass dadurch Nachteile entstehen. Bei Abbruch der Teilnahme haben Sie das Recht, die Löschung der bis dahin gesammelten Daten zu verlangen. Da die Daten im Verlauf pseudonymisiert werden (d.h. eine Zuordnung zu Ihnen dann nicht mehr möglich ist), muss eine gewünschte Löschung der eigenen Daten direkt bei Beendigung der Teilnahme beantragt werden. Aus rechtlichen Gründen dürfen Sie nur teilnehmen, wenn Sie mindestens 18 Jahre alt sind.

Für welche Zwecke sollen personenbezogene Daten verarbeitet werden?
Zweck der Speicherung und Verarbeitung von Daten ist die wissenschaftliche Nutzung im Rahmen einer Masterarbeit und ggf. auch von Publikationen, wobei jedoch keine personenbezogenen Daten veröffentlicht werden und keine Rückschlüsse auf natürliche Personen möglich sind.

Wer ist für die Datenverarbeitung verantwortlich und an wen können sich Betroffene wenden?
Universität zu Lübeck
Prof. Dr. med. Gabriele Gillessen-Kaesbach – Präsidentin
Ratzeburger Allee 160
23562 Lübeck, Deutschland
Tel.: +49 4510 3101 1000

E-Mail: praesidentin@uni-luebeck.de
Website: www.uni-luebeck.de;
www.uni-luebeck.de/univaersitaet/datenschutz.html

Datenschutzbeauftragter der Universität zu Lübeck
x-tention Informationstechnologie GmbH
Bürgermeister-Wegele-Str. 12
86167 Augsburg, Deutschland
Tel.: +49 4510 3101 1903
E-Mail: datenschutz@uni-luebeck.de

ID: _____ Datum: _____

Welche personenbezogenen Daten werden mit welcher Rechtsgrundlage verarbeitet?
Rechtsgrundlage für die Verarbeitung personengebundener Daten ist hier insbesondere Art. 6 Abs. 1 lit. A EU-Datenschutzgrundverordnung (DSGVO).

Im Rahmen der Studie erheben wir personenbezogene Daten. Dazu gehören:

- Für Gewinnverlosung: Kontaktdaten (E-Mail-Adresse)
- Soziodemografische Daten
- Angaben in Fragebögen (es werden keine Gesundheitsdaten erhoben)

Wie werden Daten verarbeitet/gespeichert und wie wird Anonymität der Teilnehmenden gewährleistet?
Die in dieser Studie getätigten Angaben werden auf den Servern des Instituts für Multimediale und Interaktive Systeme (IMIS) der Universität zu Lübeck gespeichert und ausgewertet. Dabei wird das Programm ownCloud verwendet. Lediglich der Studienleiter Univ.-Prof. Dr. rer. nat. Thomas Franke und beteiligte Forschende des IMIS haben Zugang zu den Daten der Studie.

Die E-Mail-Adresse dient ausschließlich der Gewinnbenachrichtigung bei der optionalen Teilnahme an der Gewinnverlosung. Die angegebene E-Mail-Adresse kann anonym sein und es muss an keiner Stelle in der Studie der Name angeben werden. Die E-Mail-Adresse wird getrennt von den Umfragedaten gespeichert und kann daher nicht mit Ihnen in Verbindung gebracht werden. Nach Abschluss der Gewinnverlosung werden die E-Mail-Adressen der Teilnehmenden aus dem separaten Gewinnverlosungsdatensatz gelöscht.

Im Verlauf der Studie werden die Daten mit einer Identifikationsnummer versehen. Die Daten sind dann pseudonymisiert, d.h. dass sie nur der jeweiligen Identifikationsnummer zugeordnet sind.

Welche Rechte haben Betroffene grundsätzlich?
Die Inanspruchnahme der Rechte entfaltet nur eine Wirkung, wenn die verarbeiteten Daten eine Identifizierung einer natürlichen Person zulassen.

Einwilligung und Widerruf nach Art. 7 Abs 3 DSGVO
Die Angabe personenbezogener Daten ist freiwillig. Die Einwilligung zur Verarbeitung der personenbezogenen Daten kann jederzeit nach Art. 7 Abs. 3 DSGVO mit der Folge widerrufen werden, dass die personenbezogenen Daten der betreffenden Person nicht weiterverarbeitet werden.

Auskunftsrecht (Art. 13 DSGVO)
Betroffene haben das Recht, jederzeit Auskunft über die zu ihrer Person verarbeiteten Daten sowie die möglichen Empfänger dieser Daten verlangen zu können. Ihnen steht eine Antwort innerhalb einer Frist von einem Monat nach Eingang des Auskunftsersuchens zu.

ID: _____ Datum: _____

Recht auf Berichtigung, Löschung und Einschränkung (Art. 16 – 18 DSGVO)
Die Betroffenen können jederzeit gegenüber der Universität zu Lübeck die Berichtigung, Löschung ihrer personenbezogenen Daten bzw. die Einschränkung der Verarbeitung verlangen.

Recht auf Datenübertragbarkeit (Art. 20 DSGVO)
Betroffene können verlangen, dass der Verantwortliche ihnen ihre personenbezogenen Daten in einem maschinenlesbaren Format übermittelt. Alternativ können sie die direkte Übermittlung der von ihnen bereitgestellten personenbezogenen Daten an einen anderen Verantwortlichen verlangen, soweit dies möglich ist.

Beschwerderecht (Art. 77 DSGVO)
Betroffene Personen können sich jederzeit an den Datenschutzbeauftragten der Universität zu Lübeck sowie bei einer Beschwerde nach Art. 77 DSGVO an die zuständige Aufsichtsbehörde zum Datenschutz wenden. Die zuständige Aufsichtsbehörde ist das Unabhängige Landeszentrum für Datenschutz Schleswig-Holstein (Tel.: 0431 9881 1200, E-Mail: mail@datenschutzzentrum.de).

Hiermit bestätige ich, dass ich die Datenschutz- und Einwilligungserklärung gelesen und verstanden habe und freiwillig an der Studie teilnehmen möchte.

Ort, Datum, Unterschrift

_____

ID: _____ Datum: _____

Demografie

Alter: _____

Geschlecht: _____

Studiengang: _____

Fachsemester: _____

Kontakt (nur für Gewinnausschüttung): _____

ID: _____ Datum: _____

ATI

Im Folgenden geht es um Ihre Interaktion mit technischen Systemen. Mit “technischen Systemen” sind sowohl Apps und andere Software-Anwendungen als auch komplette digitale Geräte (z.B. Handy, Computer, Fernseher, Auto-Navigation) gemeint.

Bitte geben Sie den Grad Ihrer **Zustimmung** zu folgenden Aussagen an.	**Stimmt gar nicht**	**stimmt weitgehend nicht**	**stimmt eher nicht**	**stimmt eher**	**stimmt weitgehend**	**stimmt völlig**
Ich beschäftige mich gern genauer mit technischen Systemen						
Ich probiere gern die Funktionen neuer technischer Systeme aus						
In erster Linie beschäftige ich mich mit technischen Systemen, weil ich muss						
Wenn ich ein neues technisches System vor mir habe, probiere ich es intensiv aus						
Ich verbringe sehr gern Zeit mit dem Kennenlernen eines neuen technischen Systems						
Es genügt mir, dass ein technisches System funktioniert, mir ist es egal, wie oder warum						
Ich versuche zu verstehen, wie ein technisches System genau funktioniert						
Es genügt mir, die Grundfunktionen eines technischen Systems zu kennen						
Ich versuche, die Möglichkeiten eines technischen Systems vollständig auszunutzen.						

ID: _____ Datum: _____

SIPA

Der folgende Fragebogen bezieht sich auf Ihr Verständnis vom KI-Modell durch die Erläuterung. Mit Informationen sind dabei alle Daten gemeint, mit denen das System arbeiten kann. Mit Ergebnis ist die Klassifizierung gemeint, welche am Ende der Informationsverarbeitung des Systems berechnet wird.

Bitte geben Sie den Grad Ihrer **Zustimmung** zu folgenden Aussagen an.	**Stimmt gar nicht**	**stimmt weitgehend nicht**	**stimmt eher nicht**	**stimmt eher**	**stimmt weitgehend**	**stimmt völlig**
Es war für mich transparent, welche Informationen durch das System gesammelt wurden.						
Die Informationen, die das System erfassen konnte, waren für mich erkennbar.						
Es war verständlich für mich, wie die gesammelten Informationen zum Ergebnis geführt haben.						
Die Informationsverarbeitung des Systems war für mich nachvollziehbar.						
Mit den mir zur zugänglichen Informationen war das Ergebnis vorhersehbar für mich.						
Die Informationsverarbeitung des Systems war vorhersehbar für mich.						

ID: _____ Datum: _____

FOST

Wie bewerten Sie das KI-Modell darüber hinaus?

Bitte geben Sie den Grad Ihrer **Zustimmung** zu folgenden Aussagen an.	**Stimmt gar nicht**	**stimmt weitgehend nicht**	**stimmt eher nicht**	**stimmt eher**	**stimmt weitgehend**	**stimmt völlig**
Das KI-Modell ist verlässlich						
Das KI-Modell ist präzise						
Das KI-Modell ist nachvollziehbar						
Ich kann dem KI-Modell vertrauen						
Ich kann mich nicht auf das KI-Modell verlassen						

ID: _____ Datum: _____

FOLGENDE FRAGEBÖGEN NACH DER INTERAKTION

ID: _____ Datum: _____

SIPA

Der folgende Fragebogen bezieht sich auf Ihr Verständnis vom KI-Modell nach der Interaktion mit dem Assessment-System. Mit Informationen sind dabei alle Daten gemeint, mit denen das System arbeiten kann. Mit Ergebnis ist die Klassifizierung gemeint, welche am Ende der Informationsverarbeitung des Systems berechnet wird.

Bitte geben Sie den Grad Ihrer **Zustimmung** zu folgenden Aussagen an.	**Stimmt gar nicht**	**stimmt weitgehend nicht**	**stimmt eher nicht**	**stimmt eher**	**stimmt weitgehend**	**stimmt völlig**
Es war für mich transparent, welche Informationen durch das System gesammelt wurden.						
Die Informationen, die das System erfassen konnte, waren für mich erkennbar.						
Es war verständlich für mich, wie die gesammelten Informationen zum Ergebnis geführt haben.						
Die Informationsverarbeitung des Systems war für mich nachvollziehbar.						
Mit den mir zur zugänglichen Informationen war das Ergebnis vorhersehbar für mich.						
Die Informationsverarbeitung des Systems war vorhersehbar für mich.						

ID: _____ Datum: _____

FOST

Wie bewerten Sie das KI-Modell nach der Interaktion mit dem Assessment System darüber hinaus?

Bitte geben Sie den Grad Ihrer **Zustimmung** zu folgenden Aussagen an.	**Stimmt gar nicht**	**stimmt weitgehend nicht**	**stimmt eher nicht**	**stimmt eher**	**stimmt weitgehend**	**stimmt völlig**
Das KI-Modell ist verlässlich						
Das KI-Modell ist präzise						
Das KI-Modell ist nachvollziehbar						
Ich kann dem KI-Modell vertrauen						
Ich kann mich nicht auf das KI-Modell verlassen						

ID: _____ Datum: _____

ESS

Wie bewerten Sie die Erklärungen, welche durch das System dargeboten worden sind?

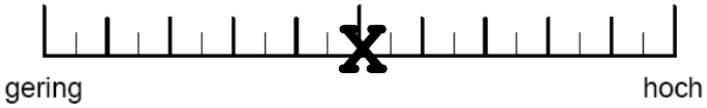
Bitte geben Sie den Grad Ihrer **Zustimmung** zu folgenden Aussagen an.	**Ich stimme voll und ganz zu**	**Ich stimme eher zu**	**Ich bin diesbezüglich neutral**	**Ich stimme eher nicht zu**	**Ich stimme überhaupt nicht zu**
Durch die Erklärung verstehe ich wie das KI-Modell funktioniert					
Diese Erklärung darüber wie das KI-Modell funktioniert ist zufriedenstellend					
Diese Erklärung darüber wie das KI-Modell funktioniert enthält genügend Details					
Diese Erklärung darüber wie das KI-Modell funktioniert erscheint vollständig zu sein					
Diese Erklärung darüber wie das KI-Modell funktioniert erklärt mir, wie ich es benutzen kann					
Diese Erklärung darüber wie das KI-Modell funktioniert ist förderlich für meine Ziele					
Diese Erklärung zeigt mir wie akkurat das KI-Modell ist					
Diese Erklärung erlaubt es mir zu beurteilen, wann ich dem KI-Modell vertrauen kann und wann nicht					

ID: _____ Datum: _____

NASA-TLX

Geben Sie jetzt für jede der unten stehenden Dimensionen an, wie hoch die Beanspruchung war. Markieren Sie dazu bitte auf den folgenden Skalen, in welchem Maße Sie sich in den sechs genannten Dimensionen von der Aufgabe beansprucht oder gefordert gesehen haben:

Beispiel:



Geistige Anforderungen
Wie viel geistige Anstrengung war bei der Informationsaufnahme und -verarbeitung erforderlich (z.B. Denken, Entscheiden, Rechnen, Erinnern, Hinsehen, Suchen...)? War die Aufgabe leicht oder anspruchsvoll, einfach oder komplex, erforderte sie hohe Genauigkeit oder war sie fehlertolerant?

Zeitliche Anforderungen
Wie viel Zeitdruck empfanden Sie hinsichtlich der Häufigkeit oder dem Takt, mit dem Aufgaben oder Aufgabenelemente auftraten? War die Abfolge langsam und geruhsam oder schnell und hektisch?

Leistung
Wie erfolgreich haben Sie Ihrer Meinung nach die vom Versuchsleiter (oder Ihnen selbst) gesetzten Ziele erreicht? Wie zufrieden waren Sie mit Ihrer Leistung bei der Verfolgung dieser Ziele?

ID: _____ Datum: _____

Anstrengung

Wie hart mussten sie arbeiten, um Ihren Grad an Aufgabenerfüllung zu erreichen?

Frustration

Wie unsicher, entmutigt, irritiert, gestresst und verärgert (versus sicher, bestätigt, zufrieden, entspannt und zufrieden mit sich selbst) fühlten Sie sich während der Aufgabe?

Assertion under Oath

I declare in lieu of an oath that I have written this paper independently and have used only the sources indicated.

[Nach Ausdruck unterschreiben. Muss auf Papier sein.]

_____

Lübeck, 27th December, 2021, Philipp Dominik Bzdok