Towards the Use of Saliency Maps for Explaining Low-Quality Electrocardiograms to End Users

Ana Lucic[1,2] , **Sheeraz Ahmad**[3] , **Amanda Furtado Brinhosa**[4] , **Q. Vera Liao**[5] ,
Himani Agrawal[6] , **Umang Bhatt**[7,8] , **Krishnaram Kenthapadi**[9] ,
Alice Xiang[10] , **Maarten de Rijke**[2] , **Nicholas Drabowski**[4] ,

[1] Partnership on AI, [2] University of Amsterdam, [3] Google, [4] Portal Telemedicina, [5] Microsoft Research,
[6] WiMLDS, [7] Mozilla Foundation, [8] University of Cambridge, [9] Fiddler AI, [10] Sony AI

Abstract

When using medical images for diagnosis, either by clinicians or artificial intelligence (AI) systems, it is important that the images are of high quality. When an image is of low quality, the medical exam that produced the image often needs to be redone. In telemedicine, a common problem is that the quality issue is only flagged once the patient has left the clinic, meaning they must return in order to have the exam redone. This can be especially difficult for people living in remote regions, who make up a substantial portion of the patients at Portal Telemedicina, a digital healthcare organization based in Brazil. In this paper, we report on ongoing work regarding (i) the development of an AI system for flagging and explaining low-quality medical images in real-time, (ii) an interview study to understand the explanation needs of stakeholders using the AI system at Portal Telemedicina, and (iii) a longitudinal user study design to examine the effect of including explanations on the workflow of the technicians in our clinics. To the best of our knowledge, this would be the first longitudinal study on evaluating the effects of XAI methods on end-users – stakeholders that use AI systems but do not have AI-specific expertise. We welcome feedback and suggestions on our experimental setup.

1 Introduction

There exist many scenarios involving AI-assisted decision making in high-stakes industries such as healthcare [Elish and Watkins, 2020; van Leeuwen *et al.*, 2021; Middleton *et al.*, 2016; Litjens *et al.*, 2017]. Explanations can help make such systems more transparent to various types of stakeholders [Mohseni *et al.*, 2018]. Prior work has found that there exists a significant gap between research and deployment for explainable AI (XAI), where current explanation techniques primarily cater to technical stakeholders rather than end users [Bhatt *et al.*, 2020]. In response, this work establishes a multistakeholder study with the goal of providing meaningful explanations to end users: individuals who interact with AI systems but do not necessarily have AI expertise themselves.

We first identify a real-world use case from Portal Telemedicina, a digital healthcare organization based in Brazil, where we believe explanations may be useful: flagging low-quality electrocardiogram (EKG) exams in real-time. Low-quality exams prevent clinicians from being able to accurately diagnose patients [Ahmed, 2011], but are often not discovered until the end of the pipeline when they are forwarded to the clinician for diagnosis. At this point, many patients have already left the clinic, meaning they must return to the clinic if it turns out the exam needs to be redone. Given that many of our patients live in remote regions of Brazil, where it can be difficult to come to a clinic in the first place, it is important to prevent patients from having to return to the clinic because of quality issues in their exams.x We hypothesize that providing explanations along with the flags will help technicians understand the issues with the EKG exams so they can ensure a correct follow-up exam in a timely manner.

In this paper, we adopt the 3-step approach recommended by Bhatt *et al.* [2020] for providing explanations to end users: (i) identifying stakeholders, (ii) engaging with each stakeholder, and (iii) understanding the purpose of the explanation. We report on work in progress on developing, deploying and evaluating an AI system for flagging and explaining low-quality medical images in real-time. We describe the outcomes of two critical studies in our development process, aimed at answering the following research questions:

RQ1: What types of explanations are most appropriate for different types of stakeholders in the context of detecting low-quality medical exams?

RQ2: How can we evaluate explanations in (a) objective terms such as a user's ability to perform a task using an explanation, and (b) subjective terms such as the impact on a user's trust in an AI system?

We answer **RQ1** by conducting an interview study with stakeholders from Portal Telemedicina to understand their explainability needs and goals, from which we design a technician-facing interface for using explaining low-quality EKGs. That is, we use our answer from **RQ1** to design a system that we plan to test in **RQ2**. We answer **RQ2** by outlining the design and procedure for a large-scale, application-grounded [Doshi-Velez and Kim, 2017], longitudinal study in order to evaluate the effect of including saliency map explanations on the workflow of technicians who perform EKG exams. We opt

for a longitudinal study setup in order to be able to (i) evaluate the system as it would exist in the real world: integrated into their regular workflow, and (ii) evaluate if including such a system results in technicians performing better EKG exams over time. This is work in progress and we hope to obtain valuable feedback on our **RQ2** design through the workshop.

2 Related Work

Our work utilizes user studies for both the design of a medical (X)AI system (**RQ1**) and the design of an evaluation of a medical (X)AI system (**RQ2**). In the following subsections, we discuss prior work related to medical AI user studies (Section 2.1) and medical XAI user studies (Section 2.2).

2.1 Medical AI User Studies

Designing AI Systems. There have been several interview studies in the context of medical AI to elicitate the needs of professional end users and design medical AI systems based on their needs. For example, Lee *et al.* [2020; 2021] design a human-AI collaborative system for stroke rehabilitation recommendation based on interviews with physical therapists. Jacobs *et al.* [2021] design an AI decision support system for antidepressant treatment selection based on semi-structured interviews with physicians.

Our work is similar to those mentioned above since it also designs an AI system for a medical task based on interviews with stakeholders involved in the development or use of the system. The main differences between these works and our work are: (i) we focus on EKGs as medical images while previous works focus on other medical tasks, (ii) our work includes an XAI component, and (iii) our work also includes a proposal for a user study to evaluate the system.

Designing and Evaluating AI Systems. Other work with a similar setup to ours is by Cai *et al.* [2019a], who create an AI system for retrieving similar historical medical images in order to aid pathologists in diagnosis. Similar to our work, their work includes user studies for both the design and evaluation of the system. Our work differs from theirs since ours includes an XAI component and a longitudinal component.

2.2 Medical XAI User Studies

The use and effectiveness of explanations in medical AI is a topic of considerable recent interest. For example, Tonekaboni *et al.* [2019] conduct an interview study with clinicians to understand their explainability needs and goals in intensive care units and emergency departments. In contrast, our work focuses on preventative medical care (i.e., medical screenings) as oppposed to acute medical care. Another distinction is that we use the findings from our interview study to implement an XAI system for end users, while the work of Tonekaboni *et al.* [2019] is more exploratory in nature. We also propose a setup for evaluating our medical XAI system. Below, we detail recent work that utilizes user studies to design and evaluate medical XAI systems.

Designing XAI Systems. There have been several works which, similar to our work, develop XAI systems based on the needs of various types of stakeholders. Cai *et al.* [2019b] conduct an interview study to understand what information

pathologists would like from an AI assistant when diagnosing prostate cancer as part of a human-AI collaborative decision making process. Xie *et al.* [2020] develop an XAI system based on the needs of physicians and radiologists for exploring chest X-rays. In contrast, we focus on a different task: detecting low-quality EKGs. Lakkaraju *et al.* [2022] interview doctors, healthcare professionals, and policymakers who already use AI explanations and find that these stakeholders prefer interactive explanations rather than static ones, specifically in the form of natural language dialogues. In contrast, we focus on static explanations because we are operating in a fairly low-resource setting and cannot accommodate the computational overhead of a sophisticated dialogue system. Unlike the works mentioned above, we also propose a user study for evaluating the effects of our XAI system.

Evaluating XAI Systems. Although there have been many user studies in the fields of medical AI, medical XAI, and XAI more broadly, we are not aware of any other studies that investigate the effect of explanations through a longitudinal study. We note that our paper is a work in progress – we propose a *design* for a user study, while the works we list below report on the *results* from their user studies.

Hegselmann *et al.* [2020] investigate if generalized additive models, which should be "inherently transparent" from an AI point of view, can be safely interpreted by doctors. Similar to our work, they design a quantitative survey with end users (in their case, clinicians) to evaluate the effectiveness of their system. This differs from our work in the medical task they focus on: predicting in-hospital mortality based on the first 48 hours of a patient's stay, as well as the absence of a longitudinal component.

Taly *et al.* [2019] evaluate saliency map explanations for diagnosing diabetic retinopathy with 10 ophthalmologists. Jin *et al.* [2021] evaluate saliency map explanations for classifying brain tumours with 1 clinician. Our user study will also evaluate saliency map explanations, but our will be longitudinal and our user study will be on a larger scale.

3 Problem Formulation

3.1 Task Description

In this work, we focus on an AI system that helps end users (i.e., nursing technicians) identify low-quality EKG exams. Low-quality exams can arise due to a variety of factors such as mistakes on the user's part (e.g., putting electrodes in incorrect locations), technical issues (e.g., fraying wires), or patient errors (e.g., moving excessively during the exam). The AI system takes as input an image of the exam and outputs whether or not the exam is of low quality. The goal of the system is to flag low-quality exams in real-time, so that the end user can redo the exam or take other remedial actions in a timely manner.

3.2 Dataset and Model

In general, the rate of low-quality medical exam across all of our clinics is approximately 7.5%. Therefore, we first collect a balanced dataset from Portal Telemedicina's proprietary database of historical EKG exams in order to train

our ML model. The dataset consists of *images* of EKG exams. We pull 10000 exams taken between 1 January 2020 to 8 September 2021, of which 5000 are low-quality. The binary low-quality label comes directly from the physicians who assess the exams: exams labelled as low-quality are unreadable by physicians. To avoid data leakage, we split the dataset into 80% training, 10% validation and 10% test based on *PatientID* (i.e., all exams from the same patient are in the same subset). Each patient has between 1 and 5 exams, with the vast majority (90%) having only 1 exam. The average age of patients in the dataset is 46 years old ($\sigma = 16.7$).

To detect low-quality exams, we use transfer learning with MobileNetV2 as the base: a convolutional neural network with inverted residual blocks and bottlenecking [Sandler *et al.*, 2018]. We train only the dense layers of MobileNetV2 using the Stochastic Gradient Descent (SGD) optimizer. We use a fixed learning rate of 0.0001 and apply batch normalization after every layer. We train our model in batch sizes of 64 on one GPU, which takes approximately 2 days. Our model has 0.97 precision and 0.44 recall on the *balanced* test set. This translates to 0.68 precision and 0.42 recall on the *unbalanced* test set. For Portal Telemedicina, this is sufficient for the first version of our system.

4 RQ1 Interview Study: Setup

To answer **RQ1**, we conduct an interview study with different types of stakeholders from Portal Telemedicina.

4.1 Study Design

We conducted 9 semi-structured interviews with participants who work at Portal Telemedicina: 2 executives, 3 developers, and 4 end users (i.e., technicians who perform medical exams). The group of participants had 3 women and 6 men. In order to understand the needs of various types of stakeholders involved in the process, the criteria for being included in the study was fairly broad: participants needed to have experience with an AI system, which could come in various forms such as development, deployment, interaction, or overseeing. Participants were recruited using internal communication tools at Portal Telemedicina. All participants completed a consent form before participating in the study and consented to being recorded during the interview.

4.2 Procedure

The interviews were conducted online and lasted approximately 60 minutes. We provided the option of having a translator present during the interviews if the participants chose to do so. All questions were asked in English, which all the participants could understand, but some made use of the translator in order to express their responses in their native language. The interviews had six components: (i) warm-up discussions, (ii) understanding the task, (iii) understanding end users, (iv) user questions and requirements, (v) feedback on XAI features, and (vi) reflecting on XAI user needs. The full interview script is available in Appendix A.

4.3 XAI Features

The main purpose of the interview study is to understand the problem space and understand which explanations work
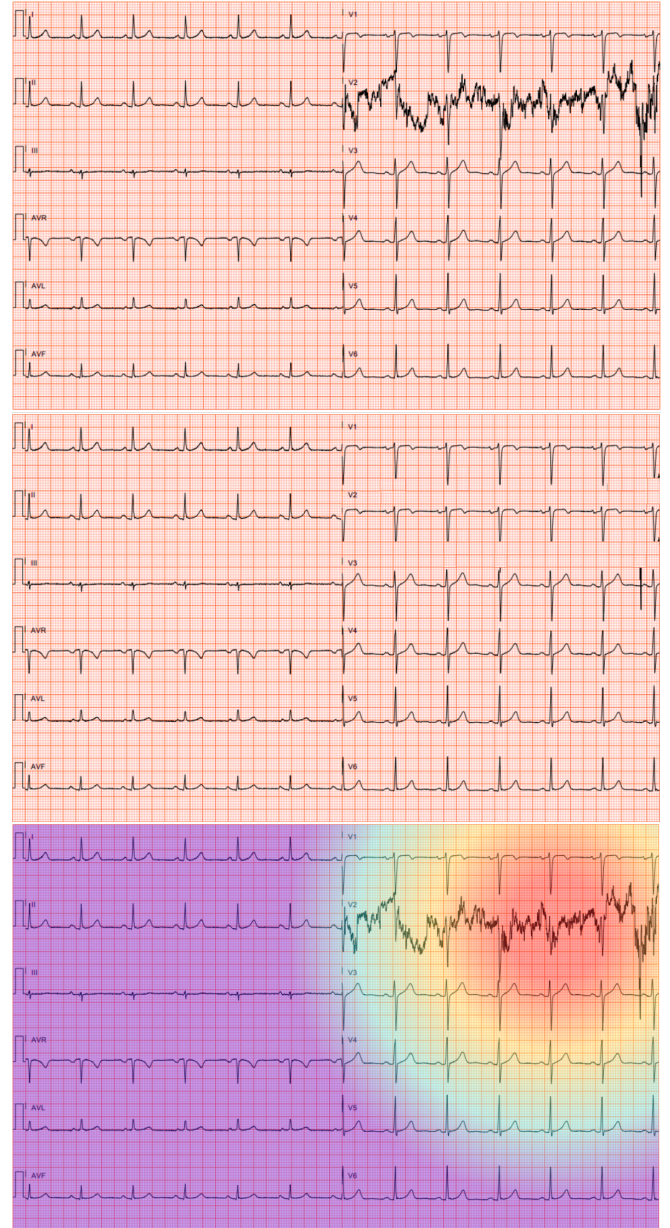

Figure 1: Behavior-based explanation probes for interview study. Top: Original example of a low-quality EKG scan. Middle: Counterfactual explanation. Bottom: Saliency map explanation for the original example.

best for which stakeholders. In the interview study, we ask participants to react to two types of explanations as defined by [Lucic *et al.*, 2021]: (i) behavior-based, and (ii) process-based. *Behavior-based* explanations provide insight into how ML models make decisions (e.g., counterfactual examples [Wachter *et al.*, 2018; Lucic *et al.*, 2022], feature attributions [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017], influential samples [Koh and Liang, 2017; Sharchilev *et al.*, 2018]). *Process-based* explanations provide insight into the processes involved in the ML modeling pipeline (e.g., model cards [Mitchell *et al.*, 2019], datasheets [Gebru *et al.*, 2018]).

Figure 1 shows the study probes we created for behavior-

based explanations. We showed users an initial example of a low-quality EKG exam (Figure 1: top). We then showed a counterfactual example (Figure 1: middle), where the problematic part of the exam is replaced, in order to show a "normal" exam. Finally, we showed a feature attribution (i.e., a saliency map) that highlights the most important part of the original image in red and the least important parts of the image in purple, with a rainbow gradient in between (Figure 1: bottom).[1] Figure 2 shows the mock model card we used as our process-based explanation probe in the interview study.

5 RQ1 Interview Study: Results

The qualitative analysis of the interviews had three stages. First, several members of our team coded the same set of three interview transcripts (one for each type of stakeholder: executive, developer, and end user – see Appendix A for details). Next, we consolidated a coherent set of themes, after which two members coded the rest of the interview transcripts according to the consolidated themes. Table 1 shows the eight main themes that emerged from the interview study outlined in Section 4.2. We group these themes into three broad categories: (i) motivation, (ii) issues, and (iii) desiderata. In the following subsections, we focus on some of the more prominent themes that came up during the interview study.

5.1 Improving Outcomes

Improving outcomes was seen by participants as one of the main motivations for including the low-quality flagging system in the pipeline. One participant broke this theme down into three main components:

"There are three benefits: benefit for the patient, because they don't need to go back to the clinic again, benefit for the clinic, there is, of course, the cost part of that, and improving the quality of the training of these technicians and the people that work with these exams."

Given that many patients are coming into the clinics from remote regions, participants felt it was important to minimize the number of patients who need to return to the clinics due to low-quality exams:

"The idea is to use AI, not only for triage, but also to detect the technical problems fast enough so that we can send these results to the clinic before the patient leaves the clinic."

5.2 Trust in the System

The degree to which stakeholders trust the AI system for low-quality exam detection was another major theme that came up during the interview process. Almost all participants touched on some aspect of this theme, especially when it came to over-trusting or under-trusting the system:

"There is this pool of people that think that AI doesn't work, and they are not open for innovation. And there are other groups of people that think that

---

Model Card: Low-Quality Exam Model

Model Details
- Convolutional neural network based on MobileNetV2 [Sandler *et al.*, 2018], implemented by Portal Telemedicina in 2021 for identifying low-quality EKG exams, input as images.

Intended Use
- Model is intended for EKG scans from machines A and B, but not machine type C.

Factors
- Gender and age group.

Metrics
- Accuracy, both over the whole population and within individual factors

Training Data
- Combination of data collected from a government database as well as scans taken at our clinics from years 2017–2018.
- Preprocessing includes mean and standard normalization.

Evaluation Data
- Same as training data, except from 2019–2020.

Ethical Considerations
- Since human lives are involved, the Brazilian Health Regulatory Agency approved the development and research of this model.

Caveats and Recommendations
- Although the model has high accuracy overall for people over 40, we do not have many data points for people 80+, so exercise caution when examining patients in this age group.

Figure 2: Process-based explanation probe for interview study: a mock model card for the low-quality exam model.

the AI will provide better success, then they leave their work to the AI – this is a problem too. We must bring both groups to the centre where they understand that the AI is trying to do a job but it's not perfect. It's an artificial intelligence, so there is an artificial 'dumbness' associated too."

5.3 Explanation Suitability

This theme emerged as a result of the questions we asked involving the interview probes shown in Figure 1 and Figure 2. Specifically, we wanted to understand which types of explanations were most useful for which stakeholders,

Table 1: Theme groupings from interview study. (<u>Underlined themes</u> are discussed further in Section 5.)

Category 1: Motivation	Category 2: Issues	Category 3: Desiderata
<u>Improving Outcomes</u>	Challenges	System Validity
Perceived Benefits of <u>XAI</u>	<u>Understanding Failures</u>	Trust in the System
Human-AI Cooperation		<u>Explanation Suitability</u>

which answers (**RQ1**). We found that the saliency maps (i.e., heatmaps) shown in Figure 1 (bottom) were the most favorably viewed explanations, across all types of stakeholders. All of the participants we interviewed found the saliency maps useful, and almost all of them believed that the saliency maps were the best option for technicians in the context of understanding why certain exams are predicted as being low-quality, including the technicians themselves. Therefore, we plan to test saliency maps explanations on our task of explaining low-quality EKG exams.

> *"I think we should only show the heat maps: the less information, the better, and the smallest part of information we can deliver here are the heatmaps."*

> *"Heatmaps are the most most friendly version of the explainability for the technicians."*

Our participants believed that counterfactual explanations, shown in Figure 1 (center), could be useful for understanding quality issues in EKG exams, especially when shown in combination with the saliency maps:

> *"It would be perfect to have them both to compare: the heat map and the counterfactual, because [the technicians] can see where the problem is with the heat map, and also an example of the correct exam with the counterfactual."*

As a result, we recommend using counterfactual explanations as a part of the training process for technicians, so they can learn to spot issues with low-quality exams by comparing them to exams that do not have quality issues. Some participants note that counterfactual explanations could also be used by clinicians in an educational context to get a better understanding of how AI systems make decisions.

The final type of explanation we tested was the mock model card shown in Figure 2. We found that most participants believed this type of explanation was best suited for stakeholders who need to have a more global view of the pipeline such as executives who make decisions about which models to productionize, or clinicians who use models to make diagnostic decisions about patients. None of our participants believed that model cards would be useful to technicians in the context of identifying low-quality exams in real-time, including the technicians themselves.

> *"I think this [model cards], this solves some questions that doctors and healthcare professionals ask. They ask how many patients we used to train, how the data was collected, they ask all these questions. I personally think this would be good for them to have these answers."*

To sum up and answer **RQ1**: our participants believe that (i) saliency maps are useful for technicians who need to un-

derstand why certain exams are flagged as low-quality in real-time, (ii) counterfactual explanations are useful as an educational tool – either for technicians during their training, or for clinicians who are using an ML model to make patient–facing decisions, and (iii) model cards are useful for stakeholders who need to have a more global view of the modeling pipeline, such as executives or clinicians.

6 RQ2 Longitudinal Study: Setup

When examining **RQ1**, we found that stakeholders from Portal Telemedicina believed saliency map explanations could be useful for explaining low-quality EKG exams to end users. To answer **RQ2**, we outline the setup for a longitudinal, application-grounded [Doshi-Velez and Kim, 2017] study to examine the effect of saliency map explanations on the workflow of technicians.

There are two components to our technician-facing system: (i) the low-quality prediction model, and (ii) the saliency map explanations. We plan to test three conditions:

- Condition A: only model prediction
- Condition B: model prediction + explanation
- Condition C: control (i.e., no input from AI system)

In our study, we will use saliency maps provided by Grad-CAM [Selvaraju *et al.*, 2016] because they are straightforward to integrate into Portal Telemedicina's pipeline. All technicians and clinics are located in Brazil and therefore this study was approved by the Brazilian Health Regulatory Agency.[2]

6.1 Study Design

In this work, we opt for a longitudinal study design as opposed to a static study design in order to understand whether or not the system is worth integrating into the technicians' workflow. A static design would only provide us with information from a single snapshot in time, whereas we want to understand the effect of including such a system on the workflow of technicians *over time*. Evaluating our system is divided into two sub-goals: (i) evaluating the technician's trust in the model prediction and its explanation, and (ii) evaluating how that translates to a lift in precision or recall of identifying low-quality exams. In order to evaluate (i), we quantify how often an exam needs to be redone following our system's prediction (and perhaps explanation) compared to the baseline of no interventions. This can depend on several subjective factors such as the perceived benefits of XAI, trust in the system, or other themes uncovered in our **RQ1** interview study (see Table 1). A high agreement with the system is indicative of the technicians' trust.

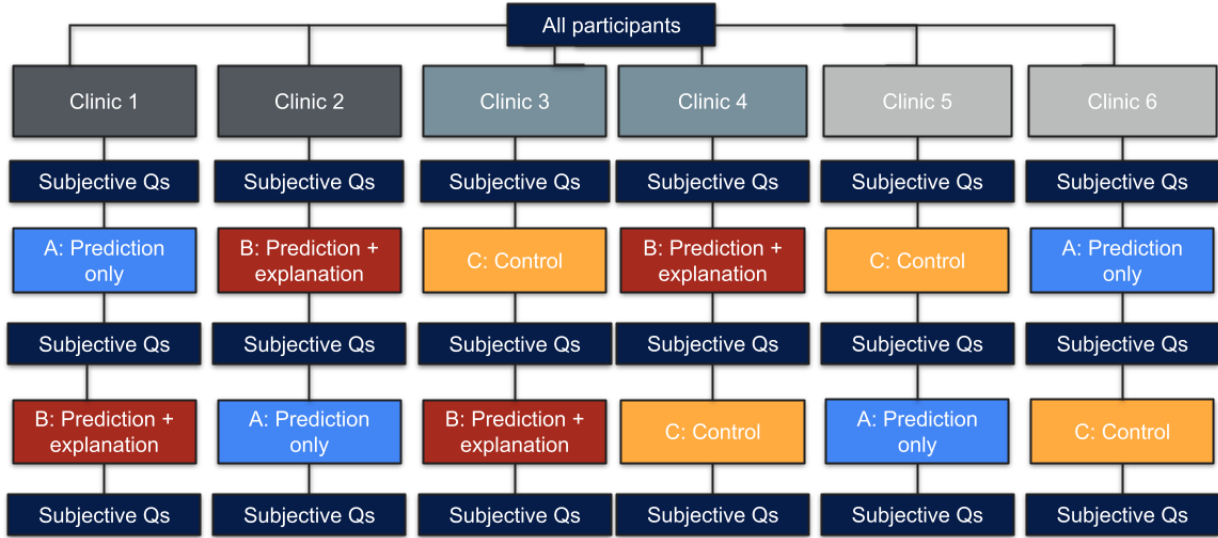[2]https://www.gov.br/anvisa/pt-br/english

Figure 3: Summary of longitudinal study setup: 6 clinics test 3 conditions (A, B, and C) in a block design.

Similarly, in order to evaluate (ii), we compare precision, the ratio of correctly redone exams to all redone exams, as well as recall, the ratio of correctly performed redone exams to all low-quality exams – across conditions A, B, and C. Improvement in precision signals better use of technician's time since fewer exams are redone unnecessarily. Improvement in recall signals better outcome for patients since the exams can be redone on the same day as opposed to a turn-around after a doctor's visit. We will obtain the ground truth labels directly from our clinicians in order to compute precision and recall. This is meant to answer **RQ2a**. To answer **RQ2b**, a subjective questionnaire is given to gauge the understanding and level of comfort with using the new system. The questionnaire is based on the Likert-scale questions proposed in [ter Hoeve *et al.*, 2017] and [Hoffman *et al.*, 2018]. See Appendix B for the full set of questions.

6.2 Procedure

Since it is not feasible to assign individual technicians to treatment and control groups, we will instead assign treatment and control conditions to entire clinics. We cannot simply assign each clinic to one condition because our clinics vary in size, the number of patients that come in, and the number of technicians that work there. Therefore, each clinic will be subjected to two different conditions and we will switch the conditions halfway through the study. We will also control for the order in which the conditions are applied, meaning we need two clinics for each pair of conditions we want to test.

Figure 3 shows a summary of the longitudinal study procedure, where 6 clinics are on-boarded for the study, conducted over 12 weeks, and all three conditions (A, B, and C) are tested in a block design. We will answer **RQ2a** by comparing results from the blue, red and yellow blocks which will be aggregated based on the condition. We will answer **RQ2b** by comparing the answers to the subjective questions, both within each clinic and aggregating over conditions.

Technicians are first given a brief introduction to machine learning, specifically how models can learn to perform classification and provide explanations in the form of saliency maps. They are also given hands-on training on how to access the new interface for accessing model's prediction and explanations, and on reaching tech support when needed. During the study, the technicians will continuously interact with the system as part of their day-to-day jobs.

For each patient, the technicians will perform the EKG exam. If the technician's clinic is under one of the treatment conditions (A or B) and the model predicts the exam is low-quality, then the technician has to make two decisions that are logged explicitly through a button in the interface: (i) do they agree with or override the model, and (ii) do they redo the exam or leave the original? Although agreeing with the model usually implies redoing the exam, there are some emergency situations where it might now, which is why we will log them separately. The **RQ2b** subjective questions will be administered three times: (i) at the beginning of the study, (ii) after the conditions switch, and (iii) at the end of the study.

7 Conclusion

In this work, we report on the ongoing development and deployment process of an AI system for detecting and explaining low-quality EKG exams. We describe the outcomes of two critical studies for shaping the development process. First, we identify which types of explanations are most appropriate for our use case by conducting a user study with stakeholders from Portal Telemedicina, in order to understand their explainability needs and goals. Second, we outline the setup for an application-grounded, longitudinal study to evaluate our system with end users. The next step in our development and deployment process is testing the effectiveness of including saliency map explanations on the workflow of technicians in our clinics and hopefully improving diagnostic outcomes for our patients.

Acknowledgements

References

[Ahmed, 2011] Hanan Ahmed. Improving diagnostic viewing of medical images using enhancement algorithms. *Journal of Computer Science*, 2011.

[Bhatt *et al.*, 2020] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *ACM Conference on Fairness, Accountability, and Transparency*, 2020.

[Cai *et al.*, 2019a] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[Cai *et al.*, 2019b] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "Hello AI": Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. In *Proceedings of the ACM on Human-Computer Interaction, CSCW*, 2019.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv*, 2017.

[Elish and Watkins, 2020] Madeleine Clare Elish and Elizabeth Anne Watkins. Repairing innovation: A study of integrating AI in clinical care. Technical report, Data & Society, 2020.

[Gebru *et al.*, 2018] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. In *ACM Conference on Fairness, Accountability, and Transparency*, 2018.

[Hegselmann *et al.*, 2020] Stefan Hegselmann, Thomas Volkert, Hendrik Ohlenburg, Antje Gottschalk, Martin Dugas, and Christian Ertmer. An evaluation of the doctor-interpretability of generalized additive models with interactions. In *Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research, 2020.

[Hoffman *et al.*, 2018] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: Challenges and prospects. *arXiv*, 2018.

[Jacobs *et al.*, 2021] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[Jin *et al.*, 2021] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. One map does not fit all: Evaluating saliency map explanation on multi-modal medical images. In *ICML Workshop on Interpretable Machine Learning in Healthcare*, 2021.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

[Lakkaraju *et al.*, 2022] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv*, 2022.

[Lee *et al.*, 2020] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. In *Proceedings of the ACM on Human-Computer Interaction, CSCW*, New York, NY, USA, 2020.

[Lee *et al.*, 2021] Min Hun Lee, Daniel P. P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez i Badia. A human-AI collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[Litjens *et al.*, 2017] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017.

[Lucic *et al.*, 2021] Ana Lucic, Madhulika Srikumar, Umang Bhatt, Alice Xiang, Ankur Taly, Q. Vera Liao, and Maarten de Rijke. A multistakeholder approach towards evaluating AI transparency mechanisms. In *CHI 2021 Workshop on Operationalizing Human-Centred Perspectives in Explainable AI*, 2021.

[Lucic *et al.*, 2022] Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. FOCUS: Flexible optimizable counterfactual explanations for tree ensembles. In *AAAI Conference on AI*, 2022.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Neural Information Processing Systems 2017*, 2017.

[Middleton *et al.*, 2016] B. Middleton, D. F. Sittig, and A. Wright. Clinical decision support: A 25 year retrospective and a 25 year vision. *Yearbook of Medical Informatics*, 2016.

[Mitchell *et al.*, 2019] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *ACM Conference on Fairness, Accountability, and Transparency*, 2019.

[Mohseni *et al.*, 2018] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv*, 2018.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[Sandler *et al.*, 2018] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *arXiv*, 2018.

[Selvaraju *et al.*, 2016] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv*, 2016.

[Sharchilev *et al.*, 2018] Boris Sharchilev, Yury Ustinovsky, Pavel Serdyukov, and Maarten de Rijke. Finding influential training samples for gradient boosted decision trees. In *International Conference on Machine Learning*, 2018.

[Taly *et al.*, 2019] Ankur Taly, Anthony Joseph, Arjun Sood, Arun Narayanaswamy, Dale Webster, David Devoud Coz, Derek Wu, Ehsan Rahimy, Greg Corrado, Jesse Smith, Jonathan Krause, Katy Blumer, Lily Peng, Michael Shumski, Naama Hammel, Rory Abbott Sayres, Scott Barb, and Zahra Rastegar. Using a deep learning algorithm and integrated gradient explanation to assist grading for diabetic retinopathy. *Ophthalmology*, 2019.

[ter Hoeve *et al.*, 2017] Maartje ter Hoeve, Mathieu Heruer, Daan Odijk, Anne Schuth, Martijn Spitters, and Maarten de Rijke. Do news consumers want explanations for personalized news rankings? In *FATREC Workshop on Responsible Recommendation*, 2017.

[Tonekaboni *et al.*, 2019] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare Conference*, 2019.

[van Leeuwen *et al.*, 2021] Kicky G. van Leeuwen, Steven Schalekamp, Matthieu J. C. M. Rutten, Bram van Ginneken, and Maarten de Rooij. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European Radiology*, 2021.

[Wachter *et al.*, 2018] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 2018.

[Xie *et al.*, 2020] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. CheXplain: Enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, April 2020.

Appendix

A RQ1 Interview Study Script

Below is the full interview script for the **RQ1** study. Some groups of questions were more applicable to certain types of stakeholders than others, which we indicate in parentheses.

1. Warm-up discussions:
   - Can you describe your role, how long you've been with the company, and what you're working on?
   - Can you describe what the low-quality scan model is?

2. Understanding the task (primarily for developers):
   - What part of the the low-quality scan model have you worked on?
   - Can you explain what type of model it is? What data is it trained on?
   - What is considered a low-quality scan? How often does it happen that a scan is not good enough and needs to be redone? What are the common reasons for this?
   - How do you usually identify a low-quality scan? What happens if it is not identified on the spot?

3. Understanding end users (primarily for developers and executives):
   - Can you describe who the target users are?
   - Have you interacted with the end users directly, or learned about them?
   - What do you believe is the main value that the low-quality scan model would deliver, or the main user problem it solves?
   - Do you foresee any challenges for the users to use the low-quality scan model?
   - What factors do you think might determine whether users would adopt or trust the low-quality scan model? Is the product team doing anything to enhance user adoption or trust?
   - Besides what we discussed, are there any other user problems or design issues the team is prioritizing to solve for the low-quality scan model?
   - What does explainability mean to you in the context of the low-quality scan model? Why does your team consider it a priority?

4. User questions and requirements:
   - Imagine you are a nurse or technician working with the low-quality scan model, what kind of questions would you ask of the system?
   - Why do you think users would want to ask that? What would a good answer look like? What would a good answer achieve?

- Are there any other questions that the system should be able to answer in order for users to use it and trust it?

5. XAI features:
   - What are some examples of XAI features that the product team has considered, or are currently developing? For each one, we ask:
     - When do you think users might need this XAI feature? How can it help the users?
     - How did the team come up with this XAI feature?
     - Do you foresee any challenges or problems with this kind of XAI feature?
   - We show users 3 examples of XAI features: saliency maps, counterfactual examples and model cards (see Figure 1 and Figure 2). For each XAI feature, we ask:
     - Do you think users might need this XAI feature? Would it help the users?
     - Do you foresee any challenges or problems with this kind of XAI feature?

6. Reflecting on XAI features and user needs:
   - Are these XAI features enough, or do you foresee any challenges that we have not covered?
   - Are there any other XAI features or information that you think the low-quality scan model could provide?
   - For developers only: What kinds of XAI features would you find useful for developing or debugging the models? Have you used any? What was your experience?

B RQ2b Subjective Questions

Below is the full list of subjective questions for **RQ2b**.

- I understand why the prediction is low-quality.
- I support using this system as a tool.
- I trust this system.
- In my opinion, this system produces mostly reasonable outputs.
- I am confident in the system. I feel that it works well.
- The outputs of the system are very predictable.
- The system is very reliable. I can count on it to be correct all the time.
- I feel that when I rely on the system, I will get the right answers.
- I am wary of the system.
- I like using the system for decision making.