User Trust on an Explainable AI-based Medical Diagnosis Support System

YAO RONG, University of Tübingen, Germany
NORA CASTNER, University of Tübingen, Germany
EFE BOZKIR, University of Tübingen, Germany
ENKELEJDA KASNECI, University of Tübingen, Germany

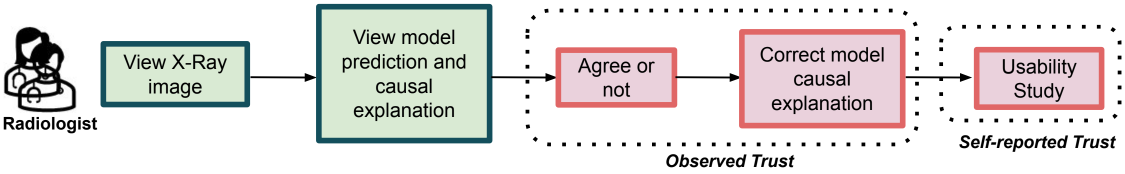


Fig. 1. **Workflow of our study:** We designed and trained an explainable AI model for Chest X-ray images. We measured the user trust and reliance based on their agreement with model explanations and final decisions (*observed trust*). Moreover, expert opinion on usability was asked to report their trust using the model and AI in general (*self-reported trust*).

Recent research has supported that system explainability improves user trust and willingness to use medical AI for diagnostic support. In this paper, we use chest disease diagnosis based on X-Ray images as a case study to investigate user trust and reliance. Building off explainability, we propose a support system where users (radiologists) can view causal explanations for final decisions. After observing these causal explanations, users provided their opinions of the model predictions and could correct explanations if they did not agree. We measured user trust as the agreement between the model's and the radiologist's diagnosis as well as the radiologists' feedback on the model explanations. Additionally, they reported their trust in the system. We tested our model on the CXR-Eye dataset and it achieved an overall accuracy of 74.1%. However, the experts in our user study agreed with the model for only 46.4% of the cases, indicating the necessity of improving the trust. The self-reported trust score was 3.2 on a scale of 1.0 to 5.0, showing that the users tended to trust the model but the trust still needs to be enhanced.

CCS Concepts: • **Human-centered computing** → **User studies**.

Additional Key Words and Phrases: medical AI support, trust, reliance, XAI

ACM Reference Format:
Yao Rong, Nora Castner, Efe Bozkir, and Enkelejda Kasneci. 2022. User Trust on an Explainable AI-based Medical Diagnosis Support System. In *Proceedings of Workshop on Trust and Reliance in AI-Human Teams at CHI 2022 (TRAIT at CHI 2022).* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

There is an ever-growing interest in employing AI-based Medical Diagnosis Support Systems (AIMDSS). However, there is often hesitation when it comes to actual integration into clinical environments [2, 6]. Medical professionals cannot afford the extra burden of a system that may react unpredictably and, more important, they may not follow its logic. When AI interaction is uncomfortable for the user, trust is diminished regarding the system's performance and practicality [2]. Recently, more effort has been directed towards improving the user trust in AIMDSS [7, 20, 33, 39]. *Trust* is defined in [21] as *"the attitude that an agent (system) will help an individual (user) achieve their goal in a situation that creates uncertainty and vulnerability for the individual"*. Other definitions of trust are more subjective: For instance, trust can be the user's faith that the model will perform well, or when the user feels comfortable working with a well-understood model [22, 25]. As eXplainable AI (XAI) techniques emerged, previous works [7, 17, 20, 25] define trust as the comfortable and confident feeling fostered when using a system. They suggested using explainable AI models to earn the trust of users. In this paper, we study user trust based on their responses to a model that presents explanations of its final diagnosis. Users could also alter these explanations, which would update the model's diagnosis.

Trust metrics used in mush of the existing research can be divided into two groups: *self-reported* and *observed* trust [26]. Self-reported trust is measured in the form of questionnaires or interviews with scores from on Likert scale, i.e., users rate how confident they feel when using the model [5, 7, 11, 20, 26–28, 30, 31, 35, 41]. Since trust includes different factors and they make *trust* more concrete and comprehensive – such as perceived reliability, perceived understandability, and faith [23] – it is measured based on multiple questions that address these different facets [20, 26]. Self-reported trust is crucial in understanding experts' feelings when interacting with the system. Observed trust is measured by observing human behavior when they use the model, which offers understanding of how they interact with the system [5, 28, 30, 31, 41]. For instance, [41, 43] used the agreement fraction (the fraction where the subject's final prediction agrees with the model's prediction) and switch fraction (the fraction where subject aligned their predictions with the model after seeing the model's prediction) to indicate human trust in a simple classification task. Observed trust metrics are necessary since the self-reported trust may not be reliable [43] and even contradictory [26]. Therefore, for more comprehensive insight to user trust, this paper is measuring trust as a combination of both self-reported and observed metrics.

The level of trust determines how *reliant* the user becomes on the system's output [21]. [28, 30, 31] also propose that explainable models help determine the appropriate level of trust. Too high of trust level can lead to *over-reliance*, where incorrect model output is overlooked due to too much faith in the system. Too low if trust level can lead to *self-reliance*, where model suggestions – even if correct – are ignored, which defeats the purpose of even having a support system. Bussone et al. [7] found increased explainability from a diagnosis decision system improved the user trust, but also led to *over-reliance*, while less explainability harmed trust, which led to *self-reliance*. In this paper, we gave users the option to change the model explanations and use the users' deviation from the model as an observed indicator for trust. By measuring these applied changes, we can determine the level of trust and how it aligns to over- and self-reliance.

The workflow of our project is illustrated in figure Figure 1. We let radiologists first view the image and then the model prediction as well as its causal explanation. Causal explanations refer to predictions of anomalies (e.g., lung opacity or lung lesion) that leads to the final diagnosis (e.g. congestive heart failure). After viewing, medical experts choose whether they agree with the model diagnosis or not, then they can change the causal explanations. Based on experts' opinions collected in the usability study, we can gain insight into the user - trust as well as their observed interaction with the model. The goal is to improve user trust and better human-AI collaboration in the context of medical decision-making by involving the expert. It is pivotal to incorporate professional feedback into the development of these systems, because ultimately these are the people who will be employing them. We want experts to *want* to use these systems.

Our contributions are as follows: We designed and trained an AI model for diagnosis based on chest X-ray images, which offers causal explanations to support its diagnoses. Then, we conducted a user study among experts to investigate the trust and usability of our model. Based on this user study, we evaluate: (1) Whether medical professionals trust our model. (2) Whether medical professionals are over-reliant or self-reliant, or have appropriate reliance. (3) How we can better design the model to increase and calibrate the trust of users.

2 RELATED WORK

XAI and user trust. Explainable AI plays an essential role in helping users to understand and appropriately trust AI models [9]. Recent research have studied how the type or the quality of explanations can affect the user trust on different tasks. For example, [26] studied a text classification model on detecting whether a Tweet was offensive or not and found that adding explanations more often harmed the user trust, especially when the explanations were low-fidelity. [28] explored the different effect on user trust when providing global and local explanations in a research-paper recommendation system. It suggested that both explanations rather than either alone could help users better understand the system. Moreover, [42] aimed to examine the effect of local explanations on the trust calibration in an income prediction task. Unfortunately, they found out that explanations did not have impact in calibrating trust. Therefore, individual user studies on trust are necessary especially in medical domain, where inappropriate trust can result in misdiagnosis [7, 20].

[7, 20] focused on how different explanation styles influenced the user trust in medical applications. Different styles refer to how the explanations are presented, for example, detailed explanations v.s. less detailed explanations in [7], or contrastive explanations in [20] where it explained as "why A and why not B". However, both user studies were conducted using simulated (fictional) data and explanations ("Wizard of Oz" approach), and only included self-reported trust measurement. Another drawback of the user study in [20] was that the participants were not from medical domain (non-experts). Compared to both previous works, our work used real-world data from a trained model and conducted the user study with the medical experts, which makes the results more convincing. Our user study also measured the observed trust as the deviation between humans' and model's explanations, as we found this suitable for the complexity of the current medical diagnosis task.

AI models for chest X-ray interpretation. Since chest X-ray images are one of the most frequent medical images in practical use, the accurate, automated analysis of chest X-ray images is becoming increasingly of interest to researchers [14, 37]. There are several very large chest X-ray datasets: MIMIC-CXR [14], Chest-Xray8 (NIH) [37], and CheXpert [12], to name only a few. The images in these datasets are labeled with approximately thirteen diagnoses [1] that can be used to train multi-label classifiers [1, 29, 34, 40]. Recently, the eye gaze data of a radiologist when reading the X-ray images was published in CXR-Eye dataset [15]. In addition, they collected three-class clinical observations from professionals in the Emergency Department (ED prognosis). Our proposed system for chest X-ray inspection differs from conventional, black-box AI tools because our model provides the causal explanation of its decision. Inspired by [19], we used the thirteen diagnoses (anomalies) to realize an “explaining" layer for predicting a final prognosis (ED prognosis) given in the dataset CXR-Eye.

[1] Each dataset has slightly different labels

3 METHODOLOGY

Our proposed work aims to find out whether users trust our explainable model and assesses their level of trust. We introduce the model's design, the causal explanations, and how it was trained in Section 3.1. Then, the user study design for the trust assessment, data collection, and analyses are detailed in Section 3.2.

3.1 Explainable Model

The system is formalized as follows: Given a chest X-ray image input $X \in \mathcal{R}^{3\times H\times W}$, an encoder $\mathcal{E}(\cdot)$ encodes the input image X into the feature vector v. The explainable layer is denoted as $f_e(\cdot)$, the final prediction (ED prognosis) layer as $f_c(\cdot)$. First, the explainable layer gives the probability of each anomaly $\hat{a} = f_e(v)$, i.e. causal explanations. Then, the final prediction layer takes $\hat{a}$ as input and classify the image as $\hat{y} = f_c(\hat{a}) = f_c(f_a(v))$. We train the explainable layer $f_e(\cdot)$ and the final prediction $f_c(\cdot)$ jointly with the following loss:

$$\begin{aligned} L_{dia} &= \beta \cdot L_e(\hat{a}, a) + L_c(\hat{y}, y) \\ &= -\beta \cdot \frac{1}{N}\sum_{i=1}^{N} a_i \cdot log(\hat{a}_i) + (1 - a_i) \cdot (1 - log(\hat{a}_i)) - \frac{1}{M}\sum_{i=1}^{M} y_i \cdot log(\hat{y}_i) \end{aligned} \tag{1}$$

Where $L_e(\cdot)$ denotes the loss of the explainable layer while $L_c(\cdot)$ the loss of the ED prognosis prediction. β is a factor that balances between the two component losses. a refers to the ground-truth causal explanation annotation for a given image and $\hat{a}$ is the prediction by $f_e(\cdot)$. N represents the amount of the causal anomalies. Since it is a multi-label classification task, $L_e(\cdot)$ calculates the binary cross-entropy loss of the prediction and the ground-truth. Similarly, y is the ground-truth label of ED prognosis and $\hat{y}$ is the predicted label using $f_c(\cdot)$. $L_c(\cdot)$ represents the cross-entropy loss between y and $\hat{y}$. M denotes the number of classes.

We trained our model on the eye gaze enhanced chest X-ray image dataset (CXR-Eye) [15]. It contains 1083 chest X-ray images from the MIMIC-CXR dataset [14]. For each image, there is a causal explanation a and each entry in a represents an anomaly (disease), 1 denotes this disease is detected in this image while 0 not, and -1 means not certain. To simplify the question, -1 is also replaced by 0 [14]. There are in total 13 anomalies. These anomalies are extracted and labelled automatically using the tool CheXpert [12]. y is the class out of the three: pneumonia, congestive heart failure (CHF) and normal, which is the ED prognosis [13].

Concretely, we used EfficientNet-b0 [36] as the backbone of the encoder $\mathcal{E}(\cdot)$. $f_e(\cdot)$ contains one fully-connected (fc) and one activation layer, while $f_c(\cdot)$ contains four fc layers. We trained the model using an Adam optimizer for max. 100 epochs and the input data X was resized to $\mathcal{R}^{3\times224\times224}$. To evaluate the model performance, we report the classification accuracy of the ED prognosis classifier and the Area Under Curve (AUC) of the causal explanations. Since the dataset has a very limited number of data, we ran a 5-fold cross-validation on our proposed model and use the averaged score from five runs as the final score. We chose $\beta = 1$, since it achieved the best performance with 74.12% accuracy and 0.67 AUC score.

3.2 User Study Design

Procedure. Our user study was designed in close collaboration with medical professionals from University Hospital Tübingen and conducted as an online questionnaire [2]. Images and predictions used in the user study were from the test set of the model. Each participant was asked to answer the questions regarding the model performance: First, they

[2] The survey is at http://hci-projects.informatik.uni-tuebingen.de/chiws.

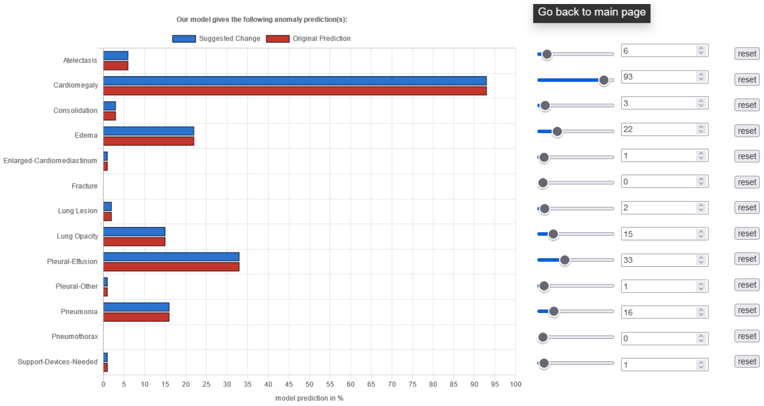


Fig. 2. User interface for investigating the model's causal explanations. Users could alter the probability of the each causal explanation.

chose whether they agreed with the model ED prognosis (pneumonia, CHF, or normal). Then, they viewed the causal explanations and could use a slider to change the probabilities of each of the 13 causal anomalies. These changes were shown as a red bar next to the model's original predictions in blue. There was an additional box for comments, where participants had the option to elaborate on their opinions (see Figure 2).

After viewing all images, they rated scores using a 5-point Likert scale ranging from one ("strongly disagree") to five ("strongly agree") on five statements in Table 1 as self-reported trust measurement. We designed our statements according to the different factors of trust, some of which also give insights into the usability of the model such as personal attachment. Two experts from the medical domain helped us adapt the questions in order to find out experts' opinions of our model performance and function implementation.

Hypothesis. We developed one main hypothesis on the expert trust in AI. In particular, we expect that users trust our model for their decision tasks on disease prediction and causal explanations. Our model was considerably accurate in diagnosis prediction on the test set with 74% accuracy (chance level being 33%). We did not divulge this performance information to the radiologists, but we anticipated that they converge on similar decisions as with the model's predictions (i.e., appropriate trust level).

Factor	Description	Question
perceived understandability	user can understand the model and predict its future behavior.	No need to learn a lot of things before getting going with this AI model.
perceived technical competence	model performs the tasks accurately and correctly based on the input information.	This AI model functioned well.
perceived reliability	model is in the sense of repeated, consistent functioning.	The AI model feedback of the Xrays was consistent.
personal attachment	user finds the model suits their taste has a strong preference for it.	The AI model was easy to use and felt confident using it.
faith	user has faith in the future ability of the model to perform even in unseen situations.	I would like to use such a kind of AI model in future work.

Table 1. Main factors in trust [23] and corresponding questions in our user study.

	Atelectasis	Cardio-megaly	Consoli-dation	Edema	Enlarged Cardio.	Fracture	Lung Lesion	Lung Opacity	Pleural Effusion	Pleural Other	Pneumonia	Pneumoth.	Support Devices
Overall	$p = 0.150$ $r = 0.303$	**$p = 0.003$** $r = 0.617$	$p = 0.284$ $r = 0.226$	**$p = 0.038$** $r = 0.437$	$p = 0.789$ $r = 0.058$	$p = 0.365$ $r = 0.191$	$p = 0.821$ $r = 0.049$	**$p = 0.0001$** $r = 0.798$	**$p = 0.003$** $r = 0.613$	$p = 0.333$ $r = 0.204$	$p = 0.323$ $r = 0.209$	$p = 0.058$ $r = 0.398$	$p = 0.551$ $r = 0.127$
Correct Prediction	$p = 0.320$ $r = 0.255$	**$p = 0.001$** $r = 0.792$	**$p = 0.046$** $r = 0.498$	**$p = 0.013$** $r = 0.610$	$p = 0.203$ $r = 0.325$	$p = 0.055$ $r = 0.481$	$p = 0.759$ $r = 0.082$	**$p = 0.013$** $r = 0.610$	**$p = 0.004$** $r = 0.697$	$p = 0.103$ $r = 0.411$	$p = 0.147$ $r = 0.368$	$p = 0.055$ $r = 0.481$	$p = 0.759$ $r = 0.08$
Incorrect Prediction	$p = 0.426$ $r = 0.333$	$p = 0.910$ $r = 0.067$	$p = 0.426$ $r = 0.333$	$p = 0.570$ $r = 0.244$	$p = 0.426$ $r = 0.333$	$p = 0.301$ $r = 0.422$	$p = 1.000$ $r = 0.022$	**$p = 0.004$** $r = 1.000$	$p = 0.469$ $r = 0.289$	$p = 0.426$ $r = 0.333$	$p = 0.652$ $r = 0.200$	$p = 0.570$ $r = 0.244$	$p = 0.652$ $r = 0.200$

Table 2. Results of the statistical analyses. For each causal explanation, we provide p values using Wilcoxon signed-rank test [38] and effect sizes with r (rank biserial correlation [16]).

Measures and Analysis. To measure the observed trust, we used the agreement rate (the percentage of times where users agreed with the model prediction) as the first indicator. As a second indicator, we analyzed each causal explanation given by users and model to see how much they differed from each other. To do this, we applied Wilcoxon signed-rank tests [38], which is a non-parametric equivalent of paired T-test, by using a significance level of $\alpha = 0.05$. The intuition behind the two indicators is that a higher agreement rate and similar distribution imply higher trust and vice versa. For each self-reported trust factor (listed in Table 1), we calculated mean scores and their standard deviations.

A subsequent analysis was carried out to determine whether users were over-reliant or self-reliant on our model by dividing the data into two groups: (a) the model made a correct prediction and (b) the model made a wrong prediction. Inside each group, we ran Wilcoxon signed-rank tests on each causal explanation to see whether users were convinced by the explanations or not. Following the definition in [7], if the alignment in (b) was high ($p < 0.05$), users were over-reliant; If the alignment in (a) was low, users were self-reliant. If the alignment in (b) was lower than (a), the users calibrated their trust fairly.

Participants. The web survey was given to the collaborating medical professionals who administered it to their co-workers. Five professionals working in the clinics participated. Two professionals had more than ten years of experience in radiology and the others had less than five years of working experience. None of them had experience working with AI models in their daily work routine. Each participant received a link having six images to investigate. This decision was to encourage experts to participate, even for just a short time, as it is hard for medical experts from clinics to set aside large time windows.

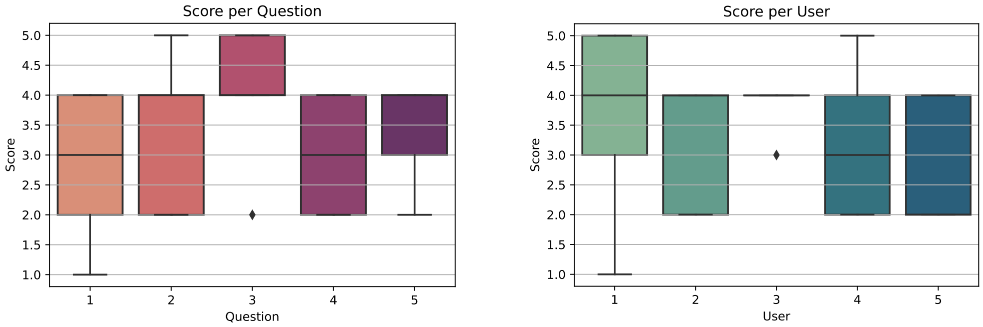
4 RESULTS

In this section, we show the results based on the data we collected and analyze whether users trust our model considering our hypothesis.

4.1 Trust

Observed Trust. To test our hypothesis, we first calculated the agreement rate between the final predictions from our model and human participants. The agreement rate was **46.4%**. We then studied the second indicator and the results are listed in Table 2. In the first row, we show the overall distribution similarity between causal explanations given by users and our model. In most cases $p > 0.05$, thus the null-hypothesis is rejected. For some of the anomalies, such as "Cardiomegaly", "Lung Opacity", and "Pleural Effusion", the p-values are either slightly below 0.05 or in the vicinity of 0.05 (e.g., for "Edema" $p \approx 0.04$). Overall, users were observed to align with our model explanations.

Self-reported Trust. In Table 1, we proposed five questions to measure trust from different perspectives. The overall trust score (in the range of 1.0 to 5.0) is calculated as the median value among five averaged trust scores (following the evaluation in [20]), which resulted in 3.2. This score indicates that users overall trusted the model, but there is still the need to improve the trust. The scores in detail are presented in Figure 3. In Figure 3a, the scores for each question regarding one trust factor are illustrated. We see that the users gave positive feedback on the factors such as "perceived reliability", "perceived understandability" and "faith in the future use" ($Mdn = 4$). However, for the "perceived technical competence" and "personal attachment", users were not very confident ($Mdn = 3$). Different users seemed to have different opinions, which is depicted in Figure 3b. User 1, 2 and 3 rated every question with higher scores, while User 5 gave more negative feedback ($Mdn = 2$), indicating that this user was not convinced by the model.



(a) Scores for each question from all users.

(b) Scores for all questions from each user.

Fig. 3. Self-reported scores for trust questions.

4.2 Reliance

To estimate which type of reliance users had, we studied user behavior when the model predicted correctly and when it did incorrectly. The results are shown in the bottom part of Table 2. Note that users were not aware of whether the model was correct or not. In all the cases except for Lung Opacity, users aligned with the explanations from wrong predictions ($p > 0.05$), which indicates that users were over-reliant on the model explanations. Users trusted the explanations for correct predictions less often, for example in Cardiomegaly, Edema, Lung Opacity and Pleural Effusion, users tended to disagree ($p < 0.05$). Moreover, the users did not calibrate their trust fairly regarding Cardiomegaly, Edema and Pleural Effusion, since they trusted more in the explanations for wrong predictions than for correct predictions.

5 DISCUSSION

The radiologists' agreement to the model rate was only **46.4%**, indicating that users had very low trust in our model, even though it had good performance on the dataset. Our model performed with an accuracy of ≈ 75% on the same images used in the user study. Users were not aware of this performance level and thus perceived it as lower performing. However, in the statistical analyses on the model explanations and the changes the radiologists made to them, we found that distributions differed only in few causal explanations. Furthermore, users tended to be over-reliant on some explanations. This result implies that users seemed to agree with most of the causal explanations provided by the model, while they were not convinced by the final prediction. These two paradoxical findings reveal that users had mixed feelings of trust and doubt in the model. The self-reported trust score, 3.2 out of 5, also validates that users leaned toward agreeing with the model predictions but not completely, suggesting they feel that the model performance should be further improved. Then, higher self-reported trust scores could be achieved, which requires further investigation both from machine learning and user-study perspective as only few misclassifications can deteriorate user trust heavily, especially in critical applications such as the medical diagnosis.

As none of the users had worked with AI models before, they were not able to calibrate their expectations or trust fairly. In this sense, model causal explanations did not have a perceivable effect on trust calibration, which is also in line with previous work [7, 43]. One possible reason for low trust could be the uncertainty in medical data, which leads to in general different opinions. For example, Schaekermann et al. [32] support that when there is uncertainty in the medical it can affect experts' perception of its output. However, uncertainty is typical in medical diagnosis, even between radiologists [3, 4, 18]. How the uncertainty of the model needs to be conveyed is highly crucial to developing user trust. Generally, humans can pick up on the level of confidence in another human, but this intuition is not as clear from an AI [2, 32]. Our work conveys the model's probability that a certain anomaly will be diagnosed and this, coupled with the priming to help train the model, led the user to disagree more with the system. Therefore, better interaction between AI and experts should promote expert control and guidance, yet clearly convey where the model's level of ability. One way of solving this issue is to train users in the medical domain to work with AI models so that they can use these models as support in their daily routines for diagnostic purposes. It may be wise to encourage medical professionals to collaborate with such systems as they would with another colleague, determining a diagnosis from a second opinion rather than blinding acceptance. [7] expressed this similar opinion towards AIMDSS: "Working in general practice is a hard job. I sit here on my own. I have to use my own knowledge. So this [system] is like having another person." Therefore, logical explanations are essential for users to understand models so that such AI models are not perceived as a black box.

However, there are several limitations that can be improved in future work. One limitation of our work is the number of experts contributing to the user study. Since medical experts are hard to acquire due to their extremely demanding schedules, a small number of experts is often the case in expertise research. One scoping review on expertise evaluated 73 sources and points out that the majority of these studies evaluate five, maybe ten experts[8]. Nevertheless, with more medical experts we could design a larger between-group user study to explore the effects of explanations in different qualities on user trust, which promotes interdisciplinary connections for better efforts towards integrating these models into professional environments. Moreover, the dataset we used for training, CXR-Eye, is rather small. In order to have a better performing model, we plan to use knowledge distillation methods [10, 24] to transfer the knowledge from a well-performing model trained on a larger chest X-Ray dataset such as MIMIC-CXR [14].

6 CONCLUSION

In this preliminary study, we designed an explainable AI model for X-Ray diagnosis, which provided causal explanations for the final prediction. We ran a user study to investigate whether the users trusted the model by collecting their perceived trust in the system as well as their agreement (observed trust) to both model prediction and explanations. In our analysis, we measured the observed trust using two indicators. The first one was the user agreement rate of the model prediction. Although our model achieved 74.12% on the test set, users did not agree with the model final prediction in 54% cases in the user study, suggesting that users did not trust the model enough. The second analysis was the Wilcoxon signed-ranked test on each causal explanation given by users and the model, where the result indicated that users were observed to align with our model explanations, even showing over-reliance on the explanations. Subsequently, perceived user trust was a self-report, where users rated the system as 3.2 out of 5, which validates that users were not confident in the model's ability.

7 ACKNOWLEDGMENTS

We thank our medical experts from University Hospital Tübingen, Asklepios Clinic St. Georg, and Birkle-Clinic in Überlingen who participated in our user studies. We acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Cluster of Excellence - Machine Learning: New Perspectives for Science, EXC number 2064/1 - Project number 390727645.

REFERENCES

[1] Sina Akbarian, Laleh Seyyed-Kalantari, Farzad Khalvati, and Elham Dolatabadi. 2020. Evaluating knowledge transfer in neural network for medical images. *arXiv preprint arXiv:2008.13574* (2020).

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[3] M Benchoufi, E Matzner-Lober, N Molinari, A-S Jannot, and P Soyer. 2020. Interobserver agreement issues in radiology. *Diagnostic and Interventional Imaging* 101, 10 (2020), 639–641.

[4] Michael A Bruno, Eric A Walker, and Hani H Abujudeh. 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics* 35, 6 (2015), 1668–1676.

[5] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[6] Samuel Budd, Emma C Robinson, and Bernhard Kainz. 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis* (2021), 102062.

[7] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[8] Andreas Gegenfurtner, Erno Lehtinen, and Roger Säljö. 2011. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational psychology review* 23, 4 (2011), 523–552.

[9] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI magazine* 40, 2 (2019), 44–58.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *stat* 1050 (2015), 9.

[11] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st international conference on intelligent user interfaces*. 164–168.

[12] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 590–597.

[13] Alistair Johnson, Bulgarelli Lucas, Pollard Tom, Horng Steven, Celi Leo Anthony, and Roger Mark. 2021. MIMIC-IV (version 1.0). *PhysioNet* (2021).

[14] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* 6, 1 (2019), 1–8.

[15] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A Krupinski, et al. 2021. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Scientific data* 8, 1 (2021), 1–18.

[16] Dave S Kerby. 2014. The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology* 3 (2014), 11–IT.

[17] Been Kim. 2015. *Interactive and interpretable machine learning models for human machine collaboration.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[18] Young W Kim and Liem T Mansfield. 2014. Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. *American Journal of Roentgenology* 202, 3 (2014), 465–470.

[19] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International Conference on Machine Learning*. PMLR, 5338–5348.

[20] Retno Larasati, Anna De Liddo, and Enrico Motta. 2020. The Effect of Explanation Styles on User's Trust.. In *ExSS-ATEC@ IUI*.

[21] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[22] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[23] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[24] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5191–5198.

[25] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[26] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).

[27] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces.* 93–100.

[28] Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popović, and Daniel S Weld. 2021. Exploring The Role of Local and Global Explanations in Recommender Systems. *arXiv preprint arXiv:2109.13301* (2021).

[29] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine* 15, 11 (2018), e1002686.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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.* 1135–1144.

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[32] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-aware ai assistants for medical data analysis. In *Proceedings of the 2020 CHI conference on human factors in computing systems.* 1–14.

[33] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel van den Bosch. 2021. Human-Centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems. *International Journal of Human-Computer Studies* (2021), 102684.

[34] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium.* World Scientific, 232–243.

[35] Matthias Söllner, Axel Hoffmann, Holger Hoffmann, and Jan Marco Leimeister. 2012. How to use behavioral research insights on trust for HCI system design. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems.* 1703–1708.

[36] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[37] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2097–2106.

[38] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics*. Springer, 196–202.

[39] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. 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.* 1–13.

[40] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. 2017. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501* (2017).

[41] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems.* 1–12.

[42] Wei Zhang, Jun Li, Zu-Bing Li, and Zhi Li. 2018. Predicting postoperative facial swelling following impacted mandibular third molars extraction by using artificial neural networks evaluation. *Scientific reports* 8, 1 (2018), 1–9.

[43] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 295–305.