

Exploring How Anomalous Model Input and Output Alerts Affect Decision-Making in Healthcare

MARISSA RADENSKY*, University of Washington, USA

DUSTIN BURSON, Microsoft, USA

RAJYA BHAIYA, Microsoft, USA

DANIEL S. WELD, University of Washington & Allen Institute for AI, USA

An important goal in the field of human-AI interaction is to help users more appropriately trust AI systems' decisions. A situation in which the user may particularly benefit from more appropriate trust is when the AI receives anomalous input or provides anomalous output. To the best of our knowledge, this is the first work towards understanding how anomaly alerts may contribute to appropriate trust of AI. In a formative mixed-methods study with 4 radiologists and 4 other physicians, we explore how AI alerts for anomalous input, very high and low confidence, and anomalous saliency-map explanations affect users' experience with mockups of an AI clinical decision support system (CDSS) for evaluating chest x-rays for pneumonia. We find evidence suggesting that the four anomaly alerts are desired by non-radiologists, and the high-confidence alerts are desired by both radiologists and non-radiologists. In a follow-up user study, we investigate how high- and low-confidence alerts affect the accuracy and thus appropriate trust of 33 radiologists working with AI CDSS mockups. We observe that these alerts do not improve users' accuracy or experience and discuss potential reasons why.

CCS Concepts: • Human-centered computing → Empirical studies in HCI; • Computing methodologies → Machine learning; • Applied computing → Health care information systems.

Additional Key Words and Phrases: human-AI interaction, explainable AI, appropriate trust of AI

1 INTRODUCTION AND RELATED WORK

Human-AI decision-making teams are present in several domains such as loan risk assessment [16, 32] and student performance forecasting [24, 73], but how to best help users appropriately trust AI models remains an open question broadly [5, 78] and in healthcare specifically [28, 37, 53, 56]. In healthcare, alert fatigue [3, 13] and high-stakes time constraints may further complicate achieving this goal. Human-AI interaction research has given significant attention to healthcare [4, 6, 10, 11, 34–36, 46, 67, 71, 76]. Though a couple recent papers challenge explainable AI's utility in healthcare [25, 30], explainable AI clinical decision support systems (CDSSs) are studied in many areas such as antibiotic treatment [43], antidepressant recommendation [37], and acute critical illness prediction [45]. One common area is chest radiography, often utilizing saliency-map explanations [2, 29, 40, 41, 44, 55, 65]. We should note, nevertheless, that concerns exist around the faithfulness and utility of saliency-map explanations [38, 61, 70]. Regarding AI's future impact on radiologists, hopes [39, 57, 68, 72] may be buttressed and concerns [51, 68] mitigated through efforts to understand and improve radiologist-AI teams. For example, Xie et al. discovered many insights around desired explanations for a chest x-ray CDSS [74]. Meanwhile, providing diagnostic advice for chest x-rays, Gaube et al. observed that radiologists more so than other physicians thought human expert suggestions were lower quality when presented as AI suggestions. They also found that incorrect suggestions led to decreased accuracy, regardless of the alleged suggester [28].

Model explanations are often sought due to perceived anomalous behavior [33]. Tomsett et al. explained that communicating uncertainty due to **anomalous model input** is important, as it helps users to construct a more accurate mental model of the AI [66]. Many works seek to detect out-of-distribution input [23, 26, 60, 62, 75], and some focus on chest x-rays in particular [7, 12, 17, 77]. Multiple works have explored how users may interact with out-of-distribution

*Work done as intern at Microsoft and PhD student at University of Washington.

data [15, 22, 48, 49, 54]. Most related to our work, Suresh et al. saw that providing information about a model’s training data did not improve user accuracy when the input image was abnormal, but participants were less likely to follow incorrect recommendations for such input [64]. With respect to **anomalous model confidence**, Suresh et al. observed that providing a model’s predicted class probabilities did not impact accuracy when the probabilities were abnormal in the sense that they were low and similar, but participants were less likely to follow incorrect recommendations with such class probabilities [64]. Also, Bussone et al. found that providing an AI CDSS’s confidence as relatively high or low only slightly impacted healthcare professionals’ trust of and reliance on the AI [9]. Regarding **anomalous model explanations**, some works have looked into determining when explanations have reduced reliability [52, 59]. In addition, DeGrave et al. saw saliency-map explanations for COVID-19 in chest radiographs that could be considered anomalous in that they depict spurious correlations developed in training [21], also seen in another work [50].

Despite all this work related to anomalous model input and output, to the best of our knowledge, this is the first work to investigate how AI anomaly alerts may help users to more appropriately trust AI. We begin with a formative study in which radiologists and other physicians are interviewed and surveyed about their reactions to anomaly-alert mockups for a CDSS used to find pneumonia in chest x-rays. We flag anomalies for three main aspects of a model’s communication with users. Two of the anomaly types are very high and low confidence with respect to the model’s confidence distribution. Another is input significantly different from the model’s training data, presented here as pediatric x-rays, which have been noted as a potential form of anomalous input for models evaluating chest radiographs [27]. The last is explanations significantly different from an expected average explanation, presented here as saliency maps focusing outside the lungs, which prior work has observed [21]. We follow up with a user study examining how the two confidence alerts impact radiologists’ accuracy in working with mockups for the same CDSS.

In summary, we make the following contributions:

- a mixed-methods formative study with 4 radiologists and 4 other physicians exploring how users of a clinical decision support system for evaluating chest x-rays react to alerts for different anomalous model input and output: very high confidence, very low confidence, anomalous input, and anomalous explanations.
- a 33-participant user study investigating high- and low-confidence alerts’ effect on radiologist-AI team accuracy.
- evidence suggesting that 1) the four proposed anomaly alerts are desired by non-radiologist physicians, and 2) high-confidence alerts are desired by both radiologists and non-radiologists, but 3) high- and low-confidence alerts are not necessarily helpful for improving radiologist-AI team accuracy.

2 STUDY 1: MIXED-METHODS FORMATIVE STUDY

2.1 Study Design

2.1.1 Research Questions. Study 1’s research questions were as follows: 1) how do users of a chest x-ray CDSS react to alerts for very high and low confidence, anomalous input, and anomalous explanations?, 2) do users think any of the alerts would help their decision-making?, and 3) how should the alerts be presented in order to be most useful?

2.1.2 Example Selection and Presentation. Each participant evaluated the same 26 chest x-rays, taken from the train set of the Kaggle RSNA Pneumonia Detection Challenge dataset [63], which has labels carefully assigned by at least one expert radiologist. The x-rays were evaluated for pneumonia using the DenseNet121 model from the TorchXRayVision library [18, 19] pre-trained on the same dataset. We used a threshold of 0.5 for determining each image’s classification, providing a high sensitivity of 95.2% at the cost of a low specificity of 57.9%. A cardiac and neuro ICU charge registered

nurse (CCRN, MSN) confirmed that all the examples were reasonably labeled for pneumonia. The saliency-map for each example was a Grad-CAM++ explanation [14] generated using publicly available PyTorch implementations [31, 47].

Each participant interacted with all 16 anomaly scenarios, consisting of every combination of anomaly type (high or low confidence, explanation, or input), alert presence (alert or no alert), and prediction accuracy (correct or incorrect). The exception was that there were two low-confidence alert scenarios with incorrect predictions and none with a correct prediction. Each participant also saw ten non-anomalous scenarios, half of which had a "pneumonia" prediction. Of the non-anomalous scenarios, one "pneumonia" and one "no-pneumonia" prediction were incorrect. Two non-anomalous scenarios were dedicated to training, and both had correct predictions, one "pneumonia" and one "no-pneumonia." To avoid too many variables for the small sample size, the example for each scenario and the prediction for each anomaly type were held constant. The prediction for each anomaly type was chosen based on when its alert would likely be more useful ("pneumonia" for high confidence and anomalous explanations and "no pneumonia" otherwise).

The 26 chest x-rays were selected from the dataset to represent each of the aforementioned scenarios. Unlike the other examples, those labeled for **anomalous input** were selected to have large black borders and appear like pediatric cases, as confirmed with a cardiac and neuro ICU charge registered nurse (CCRN, MSN). Classifying pediatric cases is a potential issue for models trained on adult x-rays. This is noted, for example, in the chest-radiograph dataset CheXpert's datasheet [27]. In order to have enough examples of incorrectly-classified anomalous input, one anomalous input that was classified correctly with 73% confidence was instead presented as classified incorrectly with 60% confidence. Furthermore, unlike the other examples, those labeled for **anomalous explanations** were selected to have a saliency-map explanation focused outside the lungs. This kind of anomalous explanation has been observed occurring in the real world as the result of spurious correlations [21]. For a confidence range of 50% to 100%, x-rays with an associated AI confidence between 53% and 55% were selected for the **anomalously-low-confidence** scenarios, and x-rays with an associated confidence between 97% and 99% were selected for the **anomalously-high-confidence** scenarios. The anomalous confidence levels were chosen to be on the extreme ends of the confidence range so that the results would generalize better to other models, whether or not their highest and lowest confidence levels tend to be as extreme. If a scenario was associated with a correct "pneumonia" prediction but not with an anomalous explanation, then the representative x-ray was selected such that its explanation mostly matched the dataset's ground-truth mask for pneumonia. To prevent bias against the AI based on initial interactions, sections of scenarios were created to prioritize showing correct predictions earlier as well as negative anomaly alerts later, as prior work shows that algorithmic aversion arises after observing an algorithm fail [24]. The examples within each section were randomly ordered.

2.1.3 Participants and Procedure. Eight participants were recruited through contacts and directories and donated their time to the study. Four are radiologists, while the other four are physicians in other areas. Two of the radiologists evaluate chest x-rays monthly, while the other participants evaluate chest x-rays weekly. The study sessions were conducted over Microsoft Teams and were around 50 minutes long. Each participant started with an initial Microsoft Forms survey of ten 7-point Likert-type questions to gauge their inclination towards using AI and familiarity with AI.¹ These questions were selected based on prior work [1, 20, 57, 58, 68, 72]. Next, the participant was shown a Microsoft PowerPoint over Microsoft Teams.² The participant was instructed on how to work with the CDSS mockups to evaluate chest x-rays for pneumonia. As recommended by Bussone et al. [9], the confidence was defined and described as going from 50% to 100%, with 50% indicating that the AI is completely unsure if the patient has pneumonia or not. The

¹The full list of questions for this survey may be found at this link.

²An example of the slides seen by a given participant may be found at this link.

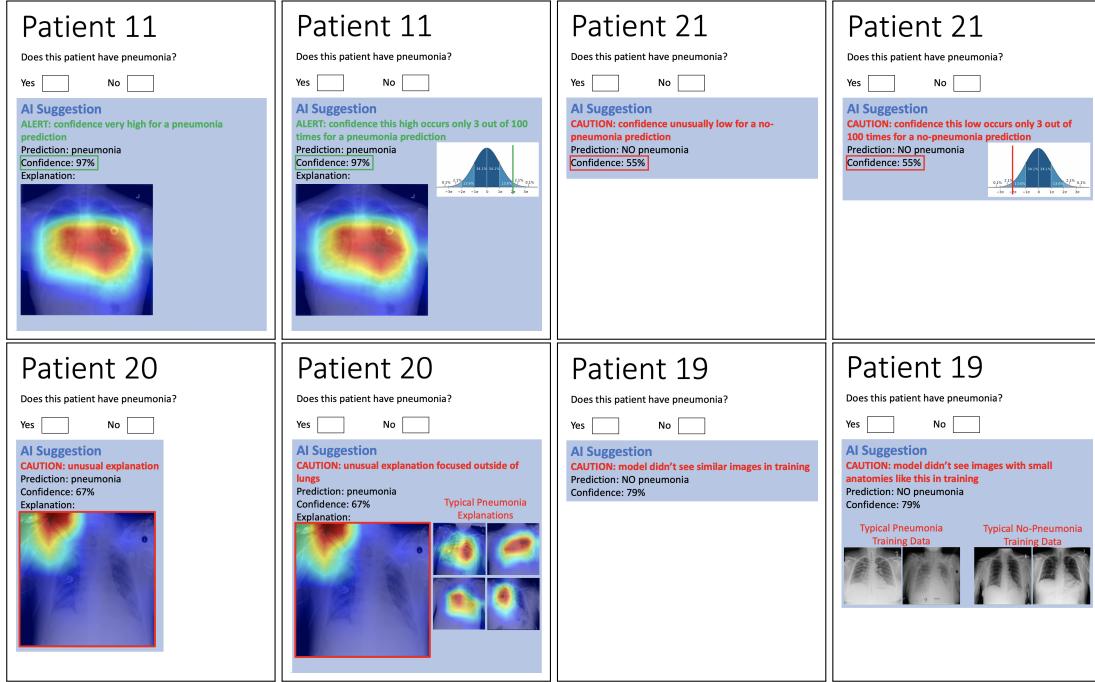


Fig. 1. Study 1 anomaly alert mockups. Top row: high confidence, high confidence detailed, low confidence, low confidence detailed. Bottom row: anomalous explanation, anomalous explanation detailed, anomalous input, anomalous input detailed.

participant then evaluated two training examples followed by 24 more. For the training examples, the researcher told them the correct answer after they made their decision. Participants were asked to think aloud [69] while evaluating the x-rays, and the researcher asked occasional questions about their interaction as well as entered their final answers.

Each x-ray was presented in a Microsoft PowerPoint slide alongside an AI suggestion, as shown in Figure 1. The prediction and confidence were always provided, and the saliency-map explanation was included if the prediction was "pneumonia." Each anomaly alert provided a short phrase describing the anomaly and highlighted the anomalous information with a colored box - green for high confidence and red otherwise. For each alert, when it was shown in a more useful scenario (i.e., when the AI was correct for the high-confidence case and incorrect for the other cases), an additional slide provided a more detailed version of the alert with a longer phrase and additional information. The simple and detailed versions of each alert are presented in Figure 1. Inspired by Xie et al. [74], the anomalous-input detailed alert included two example images of typical pneumonia training data and no-pneumonia training data, and the anomalous-explanation detailed alert included four examples of typical pneumonia saliency-maps to contrast with the anomalous one. For the high-confidence and low-confidence detailed alerts, a bell curve with a line through the point where the confidence level was served as the model's imaginary confidence distribution. This curve was not based on the model's actual confidence distribution but was merely used to see how people would react to such a visualization.

After evaluating the x-rays, the participant engaged in a semi-structured interview³ to discuss their general use of the AI, when they challenged or trusted it, when it affected their confidence, and when it affected their efficiency. These

³The full script for the study sessions may be found at this link.

questions were primarily probes for investigating how the alerts impacted their experience. They then completed a final Microsoft Forms survey consisting of two 7-point Likert-type questions for each anomaly type: 1) "The X alerts improved how I used the AI suggestions," and 2) "If I use a similar system in the future, I would like to receive X alerts." The survey reminded participants of how each alert's simple and detailed format appeared. As time permitted, the interview was recommenced to directly discuss how each alert affected their experience and any additional questions.

2.2 Results and Discussion

The think aloud and interview responses were analyzed by the first author using thematic analysis [8]. The Likert-type responses to the final survey are presented in Figure 2; please note that P7 left three questions blank.

2.2.1 High- and Low-Confidence Alerts. **The high confidence (HC) alert was desired by all but one participant, and most participants thought it improved their AI use.** This makes sense given that it is the only positive alert indicating that the AI suggestion should be more useful. However, most participants found the detailed version of the alert no more helpful than the simple one. P4 remarked, "*Confidence of 97% said the same thing to me,*" implying that the additional information was unclear in showing how *common* the confidence of 97% was. **The HC alert was primarily described as helpful for reconsidering one's initial answer when disagreeing with the AI.** Four participants saw the HC alert as useful in this way. P8 explained, "[*The HC alert*] mainly made me think more in situations where I disagreed with it." In particular, a couple participants found the HC alert useful, when disagreeing with the AI, for indicating the need to seek more information, whether in the form of another opinion or additional patient details. P2 noted, "*In that [HC alert] situation, if you disagree with the second observer, you want a third observer,*" while P3 reflected, "*I think in that situation [in which I might disagree with a HC alert], I would have then sought... more history, previous images, etc.*" On the other hand, a couple participants expressed concern that the alert might bias them towards the AI's answer. P2 commented, "*It's almost biasing me, trying to push me down the pneumonia route.*" However, P1 noted that this may not always be concerning with respect to pneumonia predictions. They explained, "*It's always better to have to overcall things and have a slightly higher false positive rate.*" Thus, HC alerts biasing users towards false-positive concern for pneumonia may be more acceptable than those biasing users towards false-negative lack of concern for pneumonia.

All the non-radiologists desired the low confidence (LC) alert, and most of them thought it improved their AI use. However, the radiologists were divided on its utility and desirability. This may be because non-radiologists rely more on the AI and thus find it more useful to recognize decreased reliability. As with the HC alerts, participants largely found the detailed version unnecessary. P2 noted, "*I think it'll confuse the average clinician.*" **The most commonly cited use for the LC alert was to recognize that one may disregard the AI suggestion.** Three radiologists described the LC alert as such. P5 commented, "*I think it made me just kind of rely on my own skills.*"

2.2.2 Anomalous-Explanation Alerts. **The anomalous-explanation (AE) alert was the least popular alert but more popular with non-radiologists.** Three non-radiologists rated the alert as desirable and having improved their AI use. As mentioned earlier, non-radiologists may like the alert more because they gain more from the AI. Regarding the detailed version, three participants acknowledged its extra words as helpful. P8 explained, "*It's clearly focused on a point outside the lung, but I think having that detail in case there was a different reason why there was something unusual is helpful.*" However, two participants noted disliking the additional words, and another two remained confused by the alert's meaning. P4 recounted, "*The words were confusing because it implied that this was some exceptional case where we actually have some brilliant conclusion.*" The alert's wording thus requires careful iteration if it is to be utilized. To

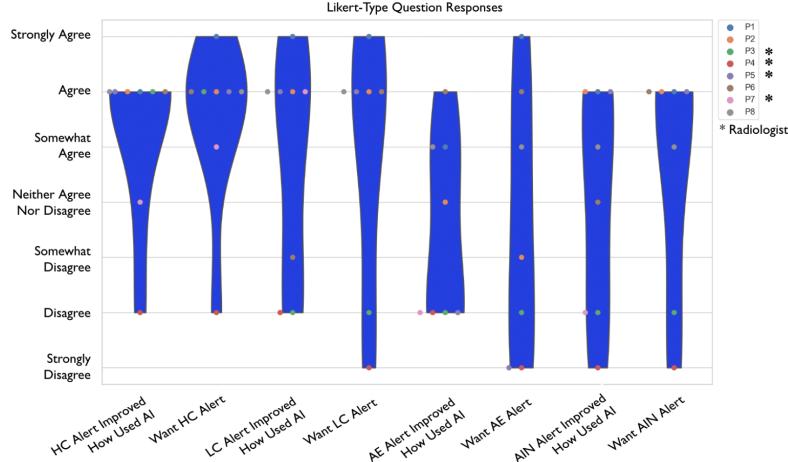


Fig. 2. Study 1 final survey responses. Color-coded dots represent individuals (see legend). Asterisks mark radiologists. We observe that HC alerts are overall desirable, and non-radiologist participants generally desire each alert. Note: P7 left Q4, Q6, and Q8 blank.

start, the alert should clarify its association with *reduced* reliability. As for the detail of showing typical pneumonia explanations, all but one participant either found it unhelpful or was confused by it.

While participants often recognized the anomalous explanations without an alert, a few participants still found the AE alert helpful for improving their understanding or trust of the AI. Participants commonly disregarded an anomalous explanation with or without the alert. For example, with no alert, P7 correctly perceived, "*The heatmap is highlighting the neck and the head? That literally makes no sense. I think there could be pneumonia for this one, but it's not where the heat map is.*" However, three non-radiologists still described the alert as helpful. For instance, P6 noted how it helped them build their understanding of the AI: "*I think that [the AE alert] also is helpful because then... you know that it's... either outside the training model, or it's picking up something that's unexpected.*" Meanwhile, the alert increased P8's trust of the AI. They said, "*It improved my confidence in the AI's abilities.... It is at least identifying that the heatmap is probably in the wrong area, and that this is not a good image or it is not well-trained in this image.*"

2.2.3 Anomalous-Input Alerts. The anomalous-input (AIN) alert was more popular with non-radiologists in terms of desirability and improving AI use. All non-radiologists rated the alert as desirable and three rated it as improving AI use. In contrast, only one radiologist viewed the AIN alert as desirable and improving AI use. As discussed previously, non-radiologists may appreciate the alert more because they rely more on the AI. Furthermore, the alert's detailed version was overall considered unhelpful. Though no participant described the typical training data as useful, three noted that it might be helpful in other contexts such as for detecting rarer diseases or for assisting junior doctors.

In response to the AIN alert, participants often ignored the AI suggestion, but participants were divided on whether or not the alert improved system use. Four participants noted disregarding the AI due to the alert. P6 explained, "*It just made me sort of discount what the AI had to say in that situation because it seemed like it was kind of outside the training model.*" Interestingly, the three most familiar with AI (P1, P2, P5) were most positive about the alert in their Likert-type responses. P2 commented, "*It's obvious the AI model hasn't seen as many kids, so that's why that caution has come up, which I think is quite useful because it's kind of telling the human... we haven't looked at as many inferences as we should do, and therefore our level of confidence is lower.*" Conversely, a couple participants found the

AIN alert unhelpful and even irritating. P3 explained, "*If AI is making an excuse for itself, then it kind of is more of a nuisance than anything.*" Likewise, P4 insisted, "*The confidence really says everything that I need to know.*" Meanwhile, P7 expressed confusion: "*I just had to read it like multiple times to figure out exactly what it was trying to say.*" Thus, the AIN alert may be improved with a recommendation to re-train the model with similar data.

2.2.4 AI Suggestions with Reduced Reliability. When asked about when the AI may have slowed them, three participants noted at least one alert. Decreasing the information provided when the AI has reduced reliability may ameliorate this issue. We asked the last six participants if they would 1) want an AI suggestion when it had reduced reliability and 2) want the AI to specify what kind of anomalous behavior (LC, AE, AIN) contributed to reduced reliability. We obtained mixed results. Two participants did not want an AI suggestion when there was reduced reliability, while the others still wanted a suggestion in at least one anomalous situation. In addition, one preferred that the AI not specify why it has reduced reliability, one wanted simply to judge the AI's reliability based on the saliency map, one was unclear, and the rest wanted specification for at least one anomaly type. Future work may investigate these questions further. Prior work indicates that AI suggestions of all confidence levels should be provided to help users calibrate their mental models of the AI [9], but systems have been designed to avoid providing AI suggestions when faced with anomalous input [17].

3 STUDY 2: USER STUDY

3.1 Study Design

3.1.1 Research Question and Hypotheses. Study 2's research question was: do alerts for very high and low AI confidence improve human-AI team accuracy in the context of a CDSS for evaluating chest x-rays? We focused on these alerts because: 1) they were most popular with Study 1's radiologists, and 2) the others seemingly required significant design iteration. Our hypotheses state that human-AI team accuracy is **H1: affected by high-confidence alerts when the AI is correct**, **H2: not affected by high-confidence alerts when the AI is incorrect**, **H3: affected by low-confidence alerts when the AI is incorrect**, and **H4: not affected by low-confidence alerts when the AI is correct**.

3.1.2 Example Labeling. Because this study was quantitative, we wanted to make sure that the example x-rays used had strongly reliable labels. To augment the dataset labels provided by at least one expert radiologist, we recruited two more radiologists with at least ten years of experience to annotate 92 selected examples and sought unanimous agreement for usable labels. After a follow-up meeting with these two labelers to discuss carefully chosen examples upon which they and the original labeler did not all agree, we still had very low agreement on labels. Suspecting that additional labelers were more likely to agree on labels for the already agreed-upon examples, we recruited two more radiologists with at least ten years of experience to label 37 examples and found that there indeed was much higher agreement. We removed the six examples upon which there was disagreement as well as two more that were paired with one of those examples for treatment assignment. Lastly, we created six synthetic examples from agreed-upon examples by changing the prediction from "pneumonia" to "no pneumonia" or vice versa. These were added to represent scenarios for which we did not otherwise have agreed-upon examples. In the end, we had 35 examples to utilize.

3.1.3 Example Selection and Presentation. The same pre-trained model [18, 19] and dataset [63] used in Study 1 were used to generate examples for Study 2. This dataset's possible labels are "pneumonia," "normal," and "lung opacity, abnormal." For simplicity, we selected images labeled "normal" rather than "lung opacity, abnormal" for our "no-pneumonia" examples, except in the case of high-confidence incorrect pneumonia predictions, for which we had no other options. After finding in Study 1 that false positives seemed more preferable than false negatives in identifying

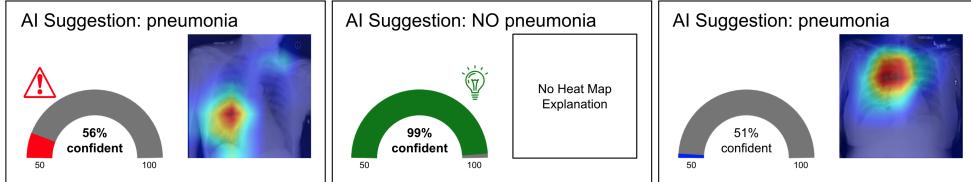


Fig. 3. Study 2 AI suggestions with a low-confidence alert, a high-confidence alert, and a baseline no-alert despite low confidence.

pneumonia, we set the threshold at 0.56 for determining each image’s classification, prioritizing sensitivity (90.2%) over specificity (70.0%) without decreasing specificity too much. The overall model accuracy was 74.5%. The saliency-map explanations were again generated as Grad-CAM++ explanations [14] with PyTorch implementations [31, 47].

Each participant encountered an alert treatment and a no-alert treatment. Per participant, there were three training examples as well as nine examples per treatment. The original confidence for each example based on the 0.56 threshold was scaled to be presented between 50% and 100%. This scaled confidence is the confidence we will refer to moving forward. The training examples as well as five examples in each treatment had a non-anomalous confidence ranging from 69% to 87%, with the exception of a training example that had a 58% confidence. The training examples consisted of one true positive, one true negative, and one false negative. The non-anomalous examples per treatment consisted of four true negatives and one true positive. Each treatment also had two low-confidence and two high-confidence examples. For each of these pairs, one was correct and one was incorrect. Also, considering the model’s confidence distribution for its training data, one had a confidence within the distribution’s fifth percentile (very-high and very-low) and the other within the twentieth percentile (high and low). The high and very-high confidences were 96% and 99% respectively. The low and very-low confidences were 51% and 56% respectively. For an anomalous example slot with a given accuracy and confidence category, the particular example was randomly chosen from the available relevant examples, which could have a prediction of “pneumonia” or “no pneumonia.” However, given the examples available, only the very-high-confidence and very-low-confidence examples could have a “pneumonia” prediction.

For each participant, each example in the alert treatment had a matching example in the no-alert treatment with respect to confidence category (normal, very-high, high, very-low, low) and prediction category (true positive, true negative, false positive, false negative). The three training examples were randomly ordered, except the incorrect one was always second. For each treatment, the anomalous examples were randomly ordered and then assigned to positions 3, 5, 7, and 8. The non-anomalous examples were also randomly ordered and filled the remaining positions. For each matching pair correctly predicting pneumonia, we made sure that the pathology masks were comparable in size.

3.1.4 Participants and Procedure. Thirty-three radiologists recruited by email from various institutions participated in Study 2 and were compensated with a \$25 Amazon gift card. All evaluated chest x-rays at least monthly (21 daily, 10 weekly, 2 monthly), and all but one had at least a year of experience in evaluating chest x-rays (<1 year: 1, 1-5 years: 13, 6-10 years: 6, >10 years: 13). Participants reported diverse sub-specialties such as cardiothoracic radiology, pediatric radiology, neuroradiology, and interventional radiology. Their demographic distribution is as follows: 21 men, 10 women, 2 unreported; 22 white, 7 Asian, 1 Native Hawaiian or Other Pacific Islander, 1 mixed race, 2 unreported; 4 Hispanic/Latinx, 25 not, 4 unreported. Also, the participants’ reported ages ranged from late twenties to late sixties.

Participants were sent a Google Forms survey⁴ and given about a week to complete it in one sitting. Three pilot runs suggested that the completion time was approximately 10 to 20 minutes. The survey began with a study description and,

⁴An example of the survey may be found at this link.

to encourage participants to do their best, noted that they would later be notified of the average participant's and their survey score. The survey then moved on to a tutorial explaining the participant's task and how to interpret the AI's output. Participants were told that they would evaluate two sets of nine chest x-rays for pneumonia, and each set would be accompanied by a different AI assistant - one with special alerts and one without. As in Study 1, we explained that a confidence of 50 indicates that the AI is completely unsure if pneumonia is present or not, while a confidence of 100 indicates the opposite. We also asked participants to assume all patients have a cough and fever, as was the assumption used in labeling the full dataset [63]. Participants were given three attention-check questions. If they did not get all three correct, their results were removed from the study. Two were removed in this way, leaving 33 participants.

Participants then moved on to evaluate 21 chest x-rays for pneumonia. The first three were training examples for which the answers were revealed to the participants. The subsequent nine comprised the first treatment and the remaining nine the second treatment. Whether the first treatment was the alert or no-alert treatment was randomized. The x-rays allocated to each treatment are described in Section 3.1.3. For each x-ray, the participants were first asked to evaluate it alone. They were provided with the x-ray itself as well as two links that allowed them to zoom in on the same x-ray and color-inverted version of it. They were asked to rate their agreement with the 7-point Likert-type statement "This patient has a lung opacity suspicious for pneumonia." This was asked instead of a binary question like in Study 1 because Study 1 participants noted that they usually provide a differential diagnosis rather than a binary answer when evaluating x-rays for pneumonia, similarly noted in Bussone et al. [9]. Next, they moved on to a page that looked identical except that it included the AI suggestion. The AI suggestion showed the AI's prediction, confidence, and saliency-map explanation if the prediction was "pneumonia" (Figure 3). Adapted from Kocielnik et al. [42], a confidence visualization was provided in Study 2 to better help people comprehend the confidence. Compared to Study 1's confidence alerts, Study 2's confidence alerts were presented in a more concise manner in order to reduce cognitive load. At the end of each treatment, participants rated their agreement with the 7-point Likert-type statement "I would use AI Assistant 1 [or 2] if it were available to me," adapted from Kocielnik et al. [42].

3.2 Results and Discussion

To evaluate each hypothesis described in Section 3.1.1, we employed the Wilcoxon signed-rank test to compare within-subjects how much closer participants got to the correct answer with versus without the alert. For example, if a participant initially gave a patient a 2 on the 7-point scale for suspicion for pneumonia and after seeing the AI suggestion gave the patient a 3, the participant got one point closer to the correct answer. Also, we considered any answer on the correct side of the Likert-type scale as correct, with 4 always incorrect. **We did not find evidence that high- or low-confidence alerts significantly impact human-AI team accuracy, whether or not the AI is correct.** The Bonferroni-corrected results did not indicate a significant difference due to alerts for correct high-confidence ($V=6$, $p=0.69$), incorrect high-confidence ($V=3$, $p=0.93$), incorrect low-confidence ($V=7.5$, $p=1.0$), or correct low-confidence ($V=10.5$, $p=1.0$) suggestions. One reason the alerts may not have had a significant effect is that the flagged confidences were very high and low in absolute terms. Had a high-confidence alert been needed for a confidence as low as 78%, for instance, participants might have gained more from a high-confidence alert. Future work may investigate if models with larger variances in their confidence distributions benefit more from anomalous-confidence alerts. Another reason may have to do with the selection of x-rays used in Study 2. The thorough process for obtaining their labels described in Section 3.1.2 only yielded examples on which four expert radiologists as well as an original expert radiologist agreed. While this process ensured that the labels were highly reliable, they were likely easier to agree upon because the associated x-rays themselves were easier to evaluate. Indeed, on average, participants provided a correct initial answer

for 7 of 8 anomalous-confidence examples. In order to better understand if confidence alerts can impact human-AI team accuracy, future work may specifically utilize data points that experts agree upon but consider difficult to evaluate. Finally, Study 2’s alerts were presented more concisely and thus perhaps more subtly than in Study 1, so they may have impacted the user less.

We used a Wilcoxon signed-rank test to compare participants’ Likert-type responses regarding whether or not they would want to use the two AI assistants. **Participants did not indicate a significant difference in preference between the AI assistant with anomalous-confidence alerts versus the one without them** ($V=24$, $p=0.44$), though the alerts assistant had a slightly higher median score (4 vs 3). This contradicts the overall positive reaction to these alerts in Study 1. That said, this negative result may have been influenced by the same points described above in relation to the hypothesis results. Furthermore, unlike in Study 1, the preference question conflated high- and low-confidence alerts; radiologists may still desire high-confidence alerts more than low-confidence ones.

4 LIMITATIONS

Given the specialized participant pool, both studies had a limited number of participants. Also, neither study was longitudinal, which would provide better insight into how the proposed interventions would work in practice. Although Study 2 provided some abilities that make evaluating x-rays easier, the studies were conducted outside of physicians’ normal environment for evaluating x-rays. Unlike in reality, participants were not provided additional information regarding each patient and were not allowed to consider diagnoses other than pneumonia. In addition, a limited number of example x-rays were used in each study and may not be representative of x-rays evaluated for pneumonia at large. Though based upon real situations, the anomalous input and explanation examples were hand-picked rather than automatically detected. Finally, Study 1’s anomalous scenarios were likely presented with unrealistically high frequency.

5 CONCLUSION

We explored how alerts for anomalous model input, confidence, and saliency-map explanations may impact users of a CDSS used to evaluate x-rays for pneumonia. In a formative study, we interviewed and surveyed 4 radiologists and 4 other physicians about their interactions with mockups of CDSS anomaly alerts. We found evidence suggesting that non-radiologist physicians who regularly evaluate chest x-rays desire the four proposed AI anomaly alerts, and high-confidence alerts are desirable among both radiologists and other physicians. In a follow-up user study, 33 radiologists engaged with two CDSS treatments, one with and one without high- and low-confidence alerts. We did not observe evidence indicating that these alerts improved radiologist-AI team accuracy or radiologists’ experience. Future work may continue to explore if and how AI suggestions should be provided in different anomalous cases. We hope that this work acts as a building block towards more research on helping users manage anomalous model input and output.

ACKNOWLEDGMENTS

This research was supported by Microsoft, NSF RAPID grant 2040196, ONR grant N00014-18-1-2193, and the Allen Institute for Artificial Intelligence (AI2). The authors thank the many people who provided helpful feedback on the project, in particular Rashmi Raj, Steven Borg, Matthew Lungren, and Ozan Oktay. The authors also thank the participants who made this work possible and the anonymous reviewers of this work for their valuable feedback.

REFERENCES

- [1] Mohamed M Abuzaid, Wiam Elshami, Huseyin Tekin, and Bashar Issa. 2022. Assessment of the willingness of radiologists and radiographers to accept the integration of artificial intelligence into radiology practice. *Academic Radiology* 29, 1 (2022), 87–94.
- [2] Md Manjurul Ahsan, Kishor Datta Gupta, Mohammad Maminur Islam, Sajib Sen, Md Rahman, Mohammad Shakhawat Hossain, et al. 2020. Study of different deep learning approach with explainable ai for screening patients with COVID-19 symptoms: Using ct scan and chest x-ray image dataset. *arXiv preprint arXiv:2007.12525* (2020).
- [3] Jessica S Ancker, Alison Edwards, Sarah Nosal, Diane Hauser, Elizabeth Mauer, and Rainu Kaushal. 2017. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC medical informatics and decision making* 17, 1 (2017), 1–9.
- [4] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. 2021. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences* 11, 11 (2021), 5088.
- [5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [6] G Octo Barnett, James J Cimino, Jon A Hupp, and Edward P Hoffer. 1987. DXplain: an evolving diagnostic decision-support system. *Jama* 258, 1 (1987), 67–74.
- [7] Christoph Berger, Magdalini Paschali, Ben Glocker, and Konstantinos Kamnitsas. 2021. Confidence-based out-of-distribution detection: a comparative study and analysis. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*. Springer, 122–132.
- [8] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012).
- [9] 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.
- [10] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.
- [11] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.
- [12] Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. 2020. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250* (2020).
- [13] Jared J Cash. 2009. Alert fatigue. *American journal of health-system pharmacy* 66, 23 (2009), 2098–2101.
- [14] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 839–847.
- [15] Changjian Chen, Jun Yuan, Yafeng Lu, Yang Liu, Hang Su, Songtao Yuan, and Shixia Liu. 2020. Oodanalyzer: Interactive analysis of out-of-distribution samples. *IEEE transactions on visualization and computer graphics* 27, 7 (2020), 3335–3349.
- [16] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.
- [17] Joseph Paul Cohen, Paul Bertin, and Vincent Frappier. 2019. Chester: A web delivered locally computed chest x-ray disease prediction system. *arXiv preprint arXiv:1901.11210* (2019).
- [18] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. 2020. On the limits of cross-domain generalization in automated X-ray prediction. In *Medical Imaging with Deep Learning*. PMLR, 136–155.
- [19] Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarnera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, et al. 2021. TorchXRayVision: A library of chest X-ray datasets and models. *arXiv preprint arXiv:2111.00595* (2021).
- [20] Fernando Collado-Mesa, Edilberto Alvarez, and Kris Arheart. 2018. The role of artificial intelligence in diagnostic radiology: a survey at a single radiology residency training program. *Journal of the American College of Radiology* 15, 12 (2018), 1753–1757.
- [21] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* 3, 7 (2021), 610–619.
- [22] Eoin Delaney, Derek Greene, and Mark T Keane. 2021. Uncertainty Estimation and Out-of-Distribution Detection for Counterfactual Explanations: Pitfalls and Solutions. *arXiv preprint arXiv:2107.09734* (2021).
- [23] Terrance DeVries and Graham W Taylor. 2018. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865* (2018).
- [24] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.
- [25] Juan Manuel Durán and Karin Rolanda Jongasma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47, 5 (2021), 329–335.
- [26] Anna Fariha, Ashish Tiwari, Arjun Radhakrishna, Sumit Gulwani, and Alexandra Meliou. 2021. Conformance constraint discovery: Measuring trust in data-driven systems. In *Proceedings of the 2021 International Conference on Management of Data*. 499–512.

- [27] Christian Garbin, Pranav Rajpurkar, Jeremy Irvin, Matthew P Lungren, and Oge Marques. 2021. Structured dataset documentation: a datasheet for chexpert. *arXiv preprint arXiv:2105.03020* (2021).
- [28] Susanne Gaube, Harini Suresh, Martina Rau, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 1–8.
- [29] Julie Gerlings, Millie Søndergaard Jensen, and Arisa Shollo. 2021. Explainable AI, but explainable to whom? *arXiv preprint arXiv:2106.05568* (2021).
- [30] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.
- [31] Jacob Gildenblat and contributors. 2021. PyTorch library for CAM methods. <https://github.com/jacobgil/pytorch-grad-cam>.
- [32] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
- [33] Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.
- [34] Diane Warner Hasling, William J Clancey, and Glenn Rennels. 1984. Strategic explanations for a diagnostic consultation system. *International Journal of Man-Machine Studies* 20, 1 (1984), 3–19.
- [35] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017).
- [36] 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. 2021. 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*. 1–14.
- [37] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.
- [38] Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).
- [39] Saurabh Jha and Eric J Topol. 2016. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *Jama* 316, 22 (2016), 2353–2354.
- [40] Md Karim, Till Döhmen, Dietrich Rebholz-Schuhmann, Stefan Decker, Michael Cochez, Oya Beyan, et al. 2020. Deepcovidexplainer: Explainable covid-19 predictions based on chest x-ray images. *arXiv preprint arXiv:2004.04582* (2020).
- [41] Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. 2021. XProtoNet: diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15719–15728.
- [42] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [43] Jean-Baptiste Lamy, Karima Sedki, and Rosy Tsopra. 2020. Explainable decision support through the learning and visualization of preferences from a formal ontology of antibiotic treatments. *Journal of Biomedical Informatics* 104 (2020), 103407.
- [44] Ricardo Bigolin Lanfredi, Ambuj Arora, Trafton Drew, Joyce D Schroeder, and Tolga Tasdizen. 2021. Comparing radiologists' gaze and saliency maps generated by interpretability methods for chest x-rays. *arXiv preprint arXiv:2112.11716* (2021).
- [45] Simon Meyer Lauritsen, Mads Kristensen, Mathias Vassard Olsen, Morten Skaarup Larsen, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, and Bo Thiesson. 2020. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature communications* 11, 1 (2020), 1–11.
- [46] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. 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*. 1–14.
- [47] WonKwang Lee. 2018. A Simple pytorch implementation of GradCAM[1], and GradCAM++[2]. https://github.com/1Konny/gradcam_plus_plus-pytorch.
- [48] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.
- [49] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. 2020. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760* (2020).
- [50] Gianluca Maguolo and Loris Nanni. 2021. A critic evaluation of methods for COVID-19 automatic detection from X-ray images. *Information Fusion* 76 (2021), 1–7.
- [51] Maciej A Mazurowski. 2019. Artificial intelligence may cause a significant disruption to the radiology workforce. *Journal of the American College of Radiology* 16, 8 (2019), 1077–1082.
- [52] Luke Merrick and Ankur Taly. 2020. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 17–38.
- [53] Massimo Micocci, Simone Borsci, Viral Thakerar, Simon Walne, Yasmine Manshadi, Finlay Edridge, Daniel Mullarkey, Peter Buckle, and George B Hanna. 2021. Do GPs Trust Artificial Intelligence Insights and What Could This Mean for Patient Care? A Case Study on GPs Skin Cancer Diagnosis in the UK. (2021).
- [54] Tshepiso Mokoena, Turgay Celik, and Vukosi Marivate. 2022. Why is this an anomaly? Explaining anomalies using sequential explanations. *Pattern Recognition* 121 (2022), 108227.

- [55] Keelin Murphy, Henk Smits, Arnoud JG Knoops, Michael BJM Korst, Tijs Samson, Ernst T Scholten, Steven Schalekamp, Cornelia M Schaefer-Prokop, Rick HHM Philipsen, Annet Meijers, et al. 2020. COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system. *Radiology* 296, 3 (2020), E166–E172.
- [56] Mohammad Naiseh. 2020. Explainability design patterns in clinical decision support systems. In *International Conference on Research Challenges in Information Science*. Springer, 613–620.
- [57] Su Kai Gideon Ooi, Andrew Makmur, Alvin Yong Quan Soon, Stephanie Fook-Chong, Charlene Liew, Soon Yiew Sia, Yong Han Ting, and Chee Yeong Lim. 2021. Attitudes toward artificial intelligence in radiology with learner needs assessment within radiology residency programmes: a national multi-programme survey. *Singapore medical journal* 62, 3 (2021), 126.
- [58] D Pinto Dos Santos, Daniel Giese, S Brodehl, SH Chon, W Staab, R Kleinert, D Maintz, and B Baeßler. 2019. Medical students' attitude towards artificial intelligence: a multicentre survey. *European radiology* 29, 4 (2019), 1640–1646.
- [59] Luyu Qiu, Yi Yang, Caleb Chen Cao, Jing Liu, Yueyuan Zheng, Hilary Hei Ting Ngai, Janet Hsiao, and Lei Chen. 2021. Resisting out-of-distribution data problem in perturbation of xai. *arXiv preprint arXiv:2107.14000* (2021).
- [60] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems* 32 (2019).
- [61] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. 2020. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence* 2, 3 (2020), e190043.
- [62] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep one-class classification. In *International conference on machine learning*. PMLR, 4393–4402.
- [63] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence* 1, 1 (2019), e180041.
- [64] Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *12th ACM Conference on Web Science*. 315–324.
- [65] Lucas O Teixeira, Rodolfo M Pereira, Diego Bertolini, Luiz S Oliveira, Loris Nanni, George DC Cavalcanti, and Yandre MG Costa. 2021. Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors* 21, 21 (2021), 7116.
- [66] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns* 1, 4 (2020), 100049.
- [67] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 359–380.
- [68] Jasper van Hoek, Adrian Huber, Alexander Leichtle, Kirsi Härmä, Daniella Hilt, Hendrik von Tengg-Kobligk, Johannes Heverhagen, and Alexander Poellinger. 2019. A survey on the future of radiology among radiologists, medical students and surgeons: students and surgeons tend to be more skeptical about artificial intelligence and radiologists may fear that other disciplines take over. *European journal of radiology* 121 (2019), 108742.
- [69] Maarten Van Someren, Yvonne F Barnard, and J Sandberg. 1994. The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress* 11 (1994).
- [70] Joseph D Viviano, Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. 2019. Saliency is a possible red herring when diagnosing poor generalization. *arXiv preprint arXiv:1910.00199* (2019).
- [71] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
- [72] Quentin Waymel, Sammy Badr, Xavier Demondion, Anne Cotten, and Thibaut Jacques. 2019. Impact of the rise of artificial intelligence in radiology: what do radiologists think? *Diagnostic and Interventional Imaging* 100, 6 (2019), 327–336.
- [73] Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A human-grounded evaluation of shap for alert processing. *arXiv preprint arXiv:1907.03324* (2019).
- [74] 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.
- [75] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2021. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334* (2021).
- [76] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [77] Jianpeng Zhang, Yutong Xie, Guansong Pang, Zhibin Liao, Johan Verjans, Wenxin Li, Zongji Sun, Jian He, Yi Li, Chunhua Shen, et al. 2020. Viral pneumonia screening on chest X-ray images using confidence-aware anomaly detection. *arXiv preprint arXiv:2003.12338* (2020).
- [78] 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.