Adjusting Doctor's Reliance on AI Through Labeling for Training Data and Modification of AI Output in a Muscle Tissue Detection Task

Keito Miyake[1,2][a], Kumi Ozaki[3][b], Akihiro Maehigashi[4][c] and Seiji Yamada[2,1][d]

[1]*Informatics Course, The Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan*
[2]*National Institute of Informatics, Tokyo, Japan*
[3]*Hamamatsu University School of Medicine, Shizuoka, Japan*
[4]*Shizuoka University, Shizuoka, Japan*

Keywords: Reliance Rate, Artificial Intelligence, Radiology, Human-AI Interaction.

Abstract: Due to the significant advancements in artificial intelligence(AI), AI technologies are increasingly providing support in various fields. However, even if AI performs at a high level, humans refuse AI for no obvious reason and prefer to solve problems on their own. For instance, experts such as medical professionals tend to be more reluctant to rely on a medical AI's diagnosis than on a human medical professional. This tendency leads to undertrust in AI and could affect its implementation in society. Thus, this study aims to mitigate the undertrust in AI by providing two functions from the perspective of interaction design: (a) labeling AI outputs as correct or incorrect for training data and (b) modifying AI outputs. To evaluate the effectiveness of these two functions in increasing medical professionals' reliance on AI, we conducted an experiment involving 25 radiologists and radiographers participating in a muscle-tissue-detection task. A two-way analysis of variance was conducted to analyze their AI-usage rate. The results indicate that both functions statistically increased reliance on AI. Our novel finding is that when radiologists are enabled to control AI output by labeling results as correct or incorrect, their reliance on AI increases.

1 INTRODUCTION

The integration of artificial intelligence(AI) into medical imaging has revolutionized the field of healthcare, particularly in areas such as radiology (Lee et al., 2023) (Sukegawa et al., 2023) (Lew et al., 2024). These advancements offer new opportunities to enhance diagnostic accuracy, improve efficiency, and ultimately provide better patient care.

As AI systems continue to demonstrate increasing capabilities in image interpretation, the reliance rate, which refers to the degree to which medical professionals rely on and depend on AI, has become an important metric for understanding how these tools are used in practice. The interaction between medical professionals and AI tools, influenced by the reliance rate, is now a critical area of study, as it affects both clinical decision-making and patient outcomes.

The rapid advancement of deep learning algorithms, particularly in the domain of computer vision, has also led to the development of AI systems capable of detecting and classifying a wide range of medical images with high accuracy.

However, research has shown that people often exhibit low reliance rates on advanced systems, even when such systems demonstrate superior performance in certain tasks (Dietvorst et al., 2018) (Logg et al., 2019). Experts in specific domains tend to exhibit higher self-efficacy compared with the general people and are less likely to use advanced systems (Gaube et al., 2021) (Jussupow et al., 2022) (Nazaretsky et al., 2022). This suggests that human errors can frequently occur more with expert than with the general population (Filiz et al., 2023) (Meyer et al., 2013).

Our motivation is to adjust the reliance rates of medical professionals on AI systems to reduce the likelihood of human errors. By optimizing this reliance, we seek to decrease the frequency of diagnostic errors that occur when AI assistance is either underused or ignored.

This paper describes our engineering approach

[a] https://orcid.org/0009-0001-3173-8011
[b] https://orcid.org/0000-0002-1454-7512
[c] https://orcid.org/0009-0000-1461-5063
[d] https://orcid.org/0000-0002-5907-7382

Miyake, K., Ozaki, K., Maehigashi, A. and Yamada, S.
Adjusting Doctor's Reliance on AI Through Labeling for Training Data and Modification of AI Output in a Muscle Tissue Detection Task.
DOI: 10.5220/0013187500003890
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 837-844
ISBN: 978-989-758-737-5; ISSN: 2184-433X

to adjust medical professionals', specifically radiologists and radiographers, reliance on AI systems, with the ultimate goal of mitigating low reliance and improving overall diagnostic performance.

Therefore, adjusting the reliance rate of medical professionals on AI systems is crucial for two reasons. First, it can lead to improved diagnostic accuracy by combining human experts' strengths with AI's computational precision. Second, it can contribute to the mitigation of human errors that occur due to low reliance on AI systems.

On the basis of these considerations, we aimed to achieve following objectives.

- investigate the effects of AI output modification and labeling on radiologists' and radiographers' reliance rates when performing diaphragm-detection tasks. This objective directly addresses the need to understand how different forms of interaction with AI systems can influence the radiologists' and radiographers' willingness to incorporate AI assistance in their diagnostic processes.
- explore how these specific AI interaction functions, modification and labeling, influence radiologists' and radiographers' decision-making processes and their perceived utility of AI assistance in diaphragm detection tasks.

By examining these objectives, we hope to contribute to developing more effective AI integration strategies to optimize the balance between human expertise and AI capabilities in clinical settings.

Previous research shows that modifying AI systems can increase medical professionals' reliance. We thus propose to provide the following two functions for increasing radiologists' reliance on such systems:

- **Labeling AI Outputs.** Users can evaluate medical images from AI, and assign labels as correct or incorrect to prepare training examples for machine learning.
- **Modification AI Outputs.** Using a digital pen, users can directly modify linear segmentation on a medical image detected by AI.

The remainder of this paper is organized as follows. Section 2 reviews related work, positioning our research within the context of literature on AI in medical imaging and strategies for adjusting reliance rates. Section 3 details our experiments, data acquisition process, and hypotheses. Section 4 presents the statistical results, focusing on comparisons between the experimental conditions. Section 5 provides a comprehensive discussion of our experimental results, and Section 6 presents our conclusions.

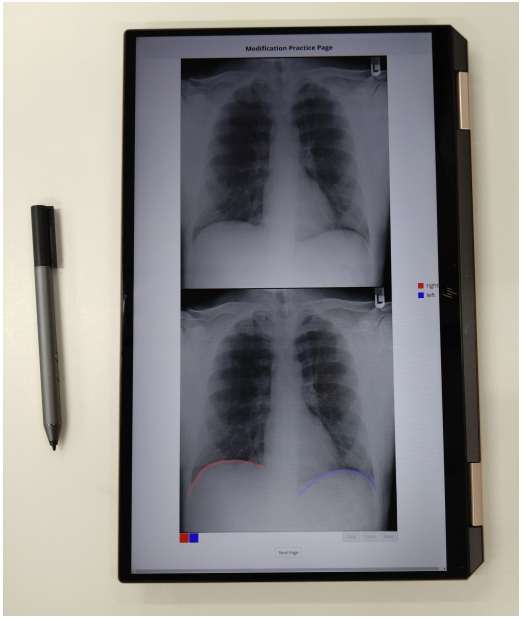

Figure 1: Touch-screen laptop and stylus used in experiments.

2 RELATED WORK

In this section, we review prior research relevant to our study, focusing on interface design for AI systems in medical settings, human responses to algorithmic forecasts, and factors affecting algorithmic reliance. We specifically examine how interface simplicity and user control influence the effectiveness of AI-assisted decision-making, particularly in medical contexts.

2.1 Factors Influencing Algorithmic Reliance in Decision-Making

(Mahmud et al., 2022) conducted an extensive systematic literature review to identify the various factors influencing reliance on algorithmic decision-making. They categorized these factors into four main themes: algorithm-related, individual, task-related, and high-level. Individual factors, such as age, experience, and familiarity with algorithms, were found to significantly influence the degree of reliance.

For instance, older individuals and those with greater professional experience tend to show lower reliance on algorithms, preferring their judgment (Arkes et al., 1986).

Conversely, less experienced individuals are more likely to rely on algorithmic outputs (Logg et al., 2019). Psychological traits and perceptions, such as trust in the algorithm and emotional responses, further contribute to the variations in reliance rates.

2.2 Impact of User Control on Reliance on AI

(Dietvorst et al., 2018) investigated the factors that influence people's willingness to use imperfect algorithms in decision-making processes. They found that giving users even the ability to modify an imperfect algorithm significantly increases their likelihood of using the algorithm.

(Cheng and Chouldechova, 2023) examined the impact of user control over algorithm design, comparing the effects of process control (enabling users to affect the input factors or algorithmic models) and outcome control (enabling users to modify algorithmic predictions). Their findings highlighted that granting users the ability to affect the training procedure significantly increased their likelihood of using the model and reduced prediction errors.

3 METHODS

3.1 Experimental Environment

The experiment was conducted in a controlled environment, with each participant completing tasks in the same room to ensure consistency in conditions.

The devices used by the participants were HP Spectre x360 laptops (display resolution 1920×1080 pixels) and the accompanying stylus. The tasks were performed in tablet mode using the stylus (Figure 2). This configuration was chosen to provide a consistent and intuitive interaction method similar to the workflow commonly used in clinical settings.

We implemented and conducted the experimental tasks using jsPsych version 7.2.1. jsPsych is a JavaScript library for psychology experiments that enables easy creation and execution of complex behavioral experiments in web browsers. The library's flexibility enabled us to precisely control the stimulus-presentation timing and record detailed response data. All participants interacted with the same web-based interface, ensuring standardized conditions across sessions. The experiment was conducted in a Chrome browser to maintain consistency in the display and response recording.

The medical images used in our experiment were sourced from the NIH Chest X-ray dataset (Wang et al., 2017). We randomly selected 40 chest X-ray images from this dataset regardless of their diagnostic difficulty. The selected images included those with clear diaphragm boundaries and those with ambiguous boundaries, providing a diverse range of image characteristics for our evaluation. Under each condition, one of the images (a total of 4 out of 40) intentionally included incorrect lines drawn by the expert to simulate errors.

3.2 Participants

Before collecting data, we conducted a priori power analysis using G*Power 3.1 (Faul et al., 2007) to determine the required sample size for our two-way repeated measures analysis of variance(ANOVA). The analysis was based on the following parameters: an effect size of 0.25, significance level of 5%, and power of 80%. The required sample size was calculated to be 24 cases.

To control for the order effects under the experimental conditions, we used a counterbalanced design. After the initial 10 baseline trials, the remaining 30 trials were presented in a randomized order across the 3 experimental conditions (modification, labeling, and combined). This randomization was implemented to minimize potential carryover effects between conditions and to control for fatigue or learning effects that might occur during the experiment. A total of 25 radiologists and radiographers with relevant experience in medical imaging (19 males, 6 females; age range: 24–54 years, median = 29; years of experience: 0-30 years, median = 3) participated in the experiment. The participants were recruited from staff members of Hamamatsu University School of Medicine Hospital. Data from all participants were included in the analysis, with no exclusions. Participants received 2,000 yen (approximately 13 USD) for completing half an hour of the experiment.

3.3 Experimental Design

The experiments involved a 2×2 within-participants design to evaluate the effectiveness of two distinct AI-interaction factors, i.e., labeling and modification. Each factor had two levels (enable vs. disable), resulting in four experimental conditions:

1. **Control Condition.** If participants chose the AI, they were unable to view the AI output and delegated the task entirely to the AI.
2. **Labeling Condition.** If participants chose the AI, they were able to label the AI output as correct or incorrect.
3. **Modification Condition.** If participants chose the AI, they were able to modify the AI output. However, if they thought the output was appropriate, they could leave it unchanged.

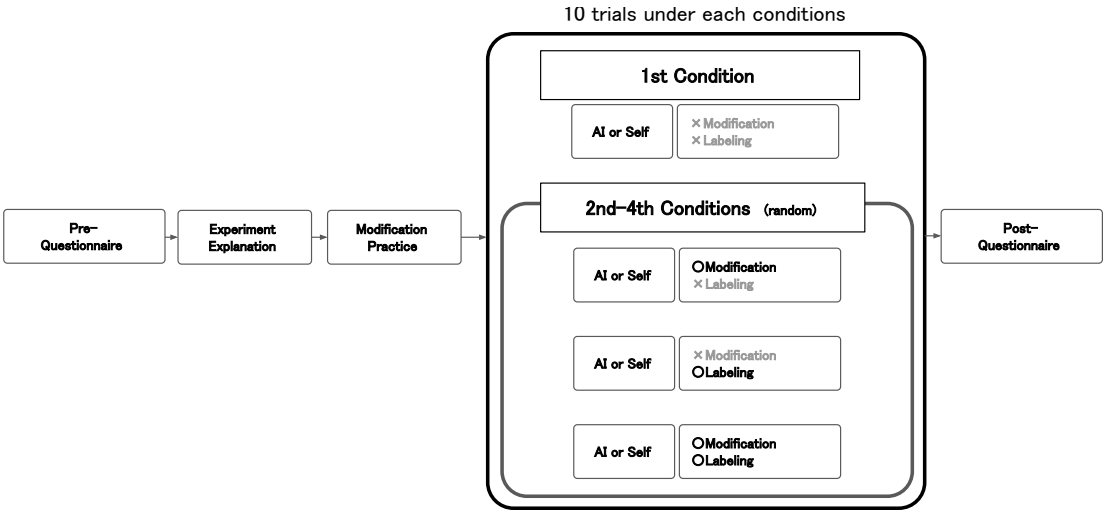

Figure 2: Overall flow of experiment.

4. **Combined Condition.** If participants chose the AI, they were able to label the AI output as correct or incorrect. If they labeled it as incorrect, they were able to modify the AI output.

3.4 Diaphragm-Detection Task

The task involved diaphragm detection on chest X-ray images (two-images are shown on the display in Figure 1). Participants were asked to detect and draw lines on diaphragm outlines on the image using a stylus interface. Specifically, participants were instructed to draw lines along the diaphragm boundaries using two colors: red for the right diaphragm and blue for the left diaphragm.

For each task trial, participants viewed an X-ray image display on a vertically oriented screen and used a stylus pen to draw the outlines. The interface enabled participants to draw precise lines and make corrections as needed. The interface also provided editing capabilities through undo, redo, and clear buttons.

3.5 Hypotheses

On the basis of the literature and identified research gap, we propose the following hypotheses:

H1. The function to label AI output increases the rate of AI utilization among medical professionals.

H2. The function to modify AI output increases the rate of AI utilization among medical professionals.

The first hypothesis is grounded in previous research in a different domain from the medical domain. Studies have shown that giving users some control over algorithmic processes can increase their willingness to use AI systems (Dietvorst et al., 2018). By extending this concept to the medical imaging domain, we aimed to test specific forms of user interaction.

The second hypothesis is motivated by the potential of user feedback to enhance the perceived reliability of AI systems. The act of labeling may engage medical professionals in a more critical evaluation of AI output, promote a sense of collaboration rather than replacement, and potentially increase their willingness to rely on AI assistance.

3.6 Procedure

Figure 2 provides an overview of the experimental process. Our experiment started with a pre-questionnaire that participants completed to gather demographic information and assess their experience with and attitudes toward AI in medical settings. The questionnaire collected data on participants' gender, age, and years of professional experience in the current position.

After the pre-questionnaire, participants were informed that the AI system was trained on data from a radiologist with over 20 years of experience. This information was provided to establish the perceived reliability and expertise of the AI system.

However, these AI outputs were directly made by the radiologist with over 20 years of experience, not generated by AI. The expert radiologist used a stylus to draw lines directly on the medical images. These pre-annotated images were presented to the participants as AI outputs.

Participants were then instructed to perform a task involving drawing red lines on the right diaphragm

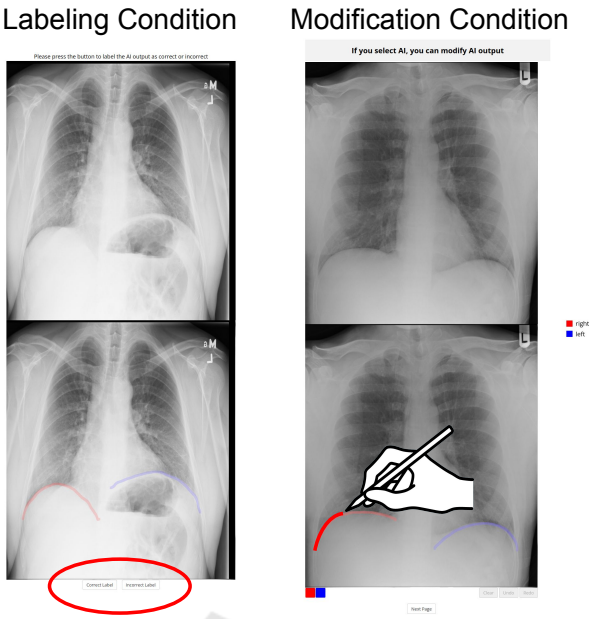

Figure 3: Task process of labeling and modification conditions.

and blue lines on the left diaphragm. For each task, they were given the option to either use AI assistance or draw the lines themselves, regardless of the experimental condition. Participants were also informed that when they chose to use AI, the specific operations they could perform with the AI output would be displayed above the original image. This information was provided to ensure participants understood what they could do and the interface for each trial.

The screen was then oriented vertically, and participants transitioned to using a stylus for input. They completed a single practice session to modify an AI output. This practice was designed to help participants become familiar with the use of the stylus and touchscreen interface, ensuring they would be comfortable with the interaction function during the main tasks.

Following the practice session, the experiment proceeded in two phases:

- Initial Reliance Rate Assessment: Participants completed ten trials with no modification and no labeling to establish an initial reliance rate.
- Main Experimental Conditions: Participants underwent 30 trials, with each trial randomly assigned to one of 3 experimental conditions: modification, labeling, or combined (both modification and labeling).

For all four conditions, each trial began with participants choosing whether to use AI or perform the task themselves. The subsequent procedure varied on the basis of this choice and the conditions (Figure 3):

1. Control Condition:
   - **If AI Was Selected.** A message indicated that the task was delegated to AI, and the trial ended.
   - **If Self-Selected.** The original X-ray image was presented for the participant to draw lines.
2. Modification Condition:
   - **If AI Was Selected.** Two images were presented, the original X-ray image (top) and the image drawn by the AI (bottom). Participants could modify the AI output or proceed without changes if they agreed with the AI.
   - **If Self-Selected.** Same as the control condition
3. Labeling Condition:
   - **If AI Was Selected.** Two images were presented as in the modification condition. Participants labeled the AI output as correct or incorrect using buttons at the bottom of the screen.
   - **If Self-Selected.** Same as the control condition.
4. Combined Condition:
   - **If AI Was Selected.** Participants first labeled the AI output as in the labeling condition. If labeled as incorrect, they could modify the AI output as in the modification condition.
   - **If Self-Selected.** Same as the control condition.

After completing all 40 trials, participants were asked to complete a post-questionnaire.

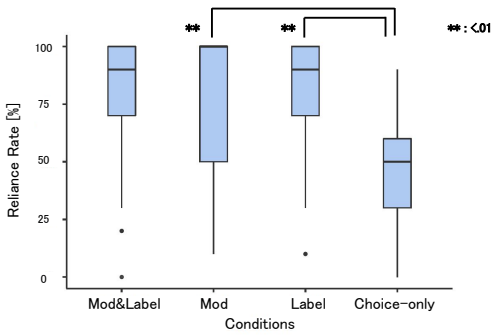

Figure 4: Box plot of reliance rates for all four conditions. Plot illustrates distribution of reliance rates under different combinations of modification and labeling functions.

4 RESULTS

We conducted a 2 (Modification: enable to modify vs. disable to modify) × 2 (Labeling: enable to label vs. disable to label) repeated measures ANOVA on reliance rate as a dependent variable. This analysis enabled us to examine the individual and combined effects of modification and labeling capabilities on participants' reliance on AI assistance in the task.

Figure 4 shows that the ANOVA revealed a significant main effects for both the modification factor ($F(1, 24) = 19.25$, $p < .001$, partial $\eta^2 = 0.45$), and labeling factor ($F(1, 24) = 24.22$, $p < .001$, partial $\eta^2 = 0.50$), as well as a significant interaction effect ($F(1, 24) = 19.39$, $p < .001$, partial $\eta^2 = 0.45$).

To further examine these effects, we investigated a simple main effect analysis using Holm's correction (Table 1). The results indicated significant simple main effects of the labeling factor under the modification factor ($F(1, 24) = 27.24$, $p_{holm} < .000$, partial $\eta^2 = 0.00$), and of the modification factor under the labeling factor ($F(1, 24) = 21.19$, $p_{holm} < .000$, partial $\eta^2 = 0.08$).

5 DISCUSSION

We aimed to investigate the influence of modification and labeling capabilities on medical professionals', specifically radiologists and radiographers, reliance on AI systems in image interpretation tasks. The results provide insights into adjusting reliance rates in medical situations. We observed the effects of both modification and labeling capabilities on reliance rates.

5.1 Hypotheses Summary

The results of our statistical analyses provide support for both hypotheses.

The first hypothesis is that the function to modify AI output will increase the rate of AI utilization among medical professionals. This hypothesis was supported by our analysis, which revealed a statistically significant increase in the reliance rate when participants were able to modify AI output compared with the control condition. This finding suggests that enabling users to modify AI output effectively encourages them to do so.

The second hypothesis proposed that the function to label AI output will increase the rate of AI utilization among medical professionals. This hypothesis was also supported. The data indicates a statistically significant increase in reliance rates when participants were able to label AI output as correct or incorrect, compared with the control condition. This indicates that the act of labeling AI output contributes to increasing acceptance and use of AI assistance.

5.2 Two Functions Effect on Reliance Rate

The function to modify AI outputs demonstrated a significant effect on reliance rates. These findings align with previous research by Dietvorst et al., who found that giving users even a small amount of modification for an imperfect algorithm significantly increases their likelihood of using it. In our medical context, the option to modify AI output likely provided participants with a sense of control over the decision. This control may have mitigated concerns about the AI system's potential errors or biases, leading to increased reliance.

The labeling functions, which enabled participants to categorize AI outputs as correct or incorrect, also significantly affected reliance rates. This finding introduces a novel perspective to the literature on human-AI interaction. The act of labeling may have several beneficial effects on user perception and behavior. Labeling encourages active engagement with AI output, prompting users to critically evaluate the system's performance. This increased engagement may lead to a better understanding of AI capabilities and limitations.

5.3 Limitation and Future Work

Though we experimentally confirmed the effectiveness of two functions to increase reliance rate, limitations should be noted. The study was conducted in

Table 1: Results of simple main effects for reliance rates across conditions.

	Conditions		Sum of Squares	df	Mean Square	F	p	η_p^2	
Mod * Label	-	Label	128.00	1	128.00	1.95	0.18	0.08	n.s.
			1572.00	24	65.50				
Mod	-	Control	12800.00	1	12800.00	21.19	0.00	0.47	$**^a$
			14500.00	24	604.17				
Mod * Label	-	Label	8.00	1	8.00	0.08	0.78	0.00	n.s.
			2492.00	24	103.83				
Label	-	Control	10952.00	1	10952.00	27.24	0.00	0.53	$**^a$
			9648.00	24	402.00				

^{a} ** indicates $p < .01$.

a controlled environment focusing on a specific muscle tissue detection task. Future research should explore these effects in diverse clinical settings, across various medical specialties, and over longer periods to understand their impact on reliance rates.

This study also focused on the increase in reliance rates, but it is important to consider this in different environments, there is a potential risk of over-reliance on AI systems (Cecil et al., 2024) (Klingbeil et al., 2024). This could lead to inappropriate decision-making if AI advice is followed uncritically. Consequently, further investigations should aim to examine methods for balancing appropriate reliance with mitigating the risk of over-reliance, ensuring that AI systems are integrated effectively without compromising human judgment.

Another critical limitation is concerned to responsibility. Since AI systems cannot take responsibility for decisions, fully relying on AI in clinical practice remains challenging. The lack of clear responsibility and accountability in AI systems means that human oversight will continue to be necessary, and this may affect the degree to which healthcare professionals are willing to rely on AI.

Due to the nature of the task, which involved drawing lines to mark specific areas, it was challenging to determine clear right or wrong answers. It was therefore difficult to compare performance outcomes between different conditions. This limitation highlights the need for further studies that incorporate more objective performance metrics to fully evaluate the impact of interaction design on user effectiveness.

Another avenue for future research involves examining the interaction between the attribution of responsibility to data providers or holders and the authority of supervising clinicians. Investigating these dynamics could offer new insights into how different accountability structures influence reliance on AI systems in healthcare settings. Understanding this interaction could help address concerns about both under-reliance and over-reliance, ensuring that AI is used as a supportive tool rather than an unquestioned authority.

6 CONCLUSIONS

We examined how modification and labeling of AI outputs affect medical professionals' reliance on AI systems in diaphragm detection tasks. Our findings indicate that both functions can significantly increase reliance rates.

The function to modify AI outputs and label them as correct or incorrect appeared to increase user reliance, potentially by providing a sense of control and encouraging critical evaluation. These results have important implications for the design of AI systems in medical settings, suggesting that incorporating the interactive functions may optimize reliance rates and improve AI integration in clinical practice. Our study will contribute to the growing research on human-AI interaction in healthcare, offering insights into how interaction design influences user behavior.

As AI systems become increasingly prevalent in clinical settings, optimizing human-AI interaction remains crucial. While our findings suggest promising avenues for improving AI integration in medical image interpretation, the complexity of balancing increased reliance with appropriate human oversight also remains to be explored. Continued research in this area is essential to ensure that AI tools enhance the quality of patient care.

ACKNOWLEDGEMENTS

This work was partially supported by JST, CREST (JPMJCR21D4), Japan.

REFERENCES

Arkes, H. R., Dawes, R. M., and Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, 37(1):93–110.

Cecil, J., Lermer, E., Hudecek, M. F. C., Sauer, J., and Gaube, S. (2024). Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task. *Scientific Reports*, 14(1):1–15.

Cheng, L. and Chouldechova, A. (2023). Overcoming algorithm aversion: A comparison between process and outcome control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, number Article 756 in CHI '23, pages 1–27.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191.

Filiz, I., Judek, J. R., Lorenz, M., and Spiwoks, M. (2023). The extent of algorithm aversion in decision-making situations with varying gravity. *PLOS One*, 18(2):e0278751.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., and Ghassemi, M. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Medcine*, 4(1):31.

Jussupow, E., Spohrer, K., and Heinzl, A. (2022). Radiologists' usage of diagnostic AI systems. *Business & Information Systems Engineering*, 64(3):293–309.

Klingbeil, A., Grützner, C., and Schreck, P. (2024). Trust and reliance on AI — an experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160:108352.

Lee, J. H., Hong, H., Nam, G., Hwang, E. J., and Park, C. M. (2023). Effect of human-AI interaction on detection of malignant lung nodules on chest radiographs. *Radiology*, 307(5):e222976.

Lew, C. O., Calabrese, E., Chen, J. V., Tang, F., Chaudhari, G., Lee, A., Faro, J., Juul, S., Mathur, A., McKinstry, R. C., Wisnowski, J. L., Rauschecker, A., Wu, Y. W., and Li, Y. (2024). Artificial intelligence outcome prediction in neonates with encephalopathy (AI-OPiNE). *Radiology: Artificial Intelligence*, page e240076.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., and Smolander, K. (2022). What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390.

Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., and Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Internal Medicine*, 173(21):1952–1958.

Nazaretsky, T., Ariely, M., Cukurova, M., and Alexandron, G. (2022). Teachers' trust in AI -powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, 53(4):914–931.

Sukegawa, S., Ono, S., Tanaka, F., Inoue, Y., Hara, T., Yoshii, K., Nakano, K., Takabatake, K., Kawai, H., Katsumitsu, S., Nakai, F., Nakai, Y., Miyazaki, R., Murakami, S., Nagatsuka, H., and Miyake, M. (2023). Effectiveness of deep learning classifiers in histopathological diagnosis of oral squamous cell carcinoma by pathologists. *Scientific Reports*, 13(1):1–9.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (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 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3462–3471. IEEE.