Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making

MIN HUN LEE and CHONG JUN CHEW, Singapore Management University, Singapore

Artificial intelligence (AI) is increasingly being considered to assist human decision-making in high-stake domains (e.g. health). However, researchers have discussed an issue that humans can over-rely on wrong suggestions of the AI model instead of achieving human AI complementary performance. In this work, we utilized salient feature explanations along with what-if, counterfactual explanations to make humans review AI suggestions more analytically to reduce overreliance on AI and explored the effect of these explanations on trust and reliance on AI during clinical decision-making. We conducted an experiment with seven therapists and ten laypersons on the task of assessing post-stroke survivors' quality of motion, and analyzed their performance, agreement level on the task, and reliance on AI without and with two types of AI explanations. Our results showed that the AI model with both salient features and counterfactual explanations assisted therapists and laypersons to improve their performance and agreement level on the task when 'right' AI outputs are presented. While both therapists and laypersons over-relied on 'wrong' AI outputs, counterfactual explanations assisted both therapists and laypersons to reduce their over-reliance on 'wrong' AI outputs by 21% compared to salient feature explanations. Specifically, laypersons had higher performance degrades by 18.0 f1-score with salient feature explanations and 14.0 f1-score with counterfactual explanations than therapists with performance degrades of 8.6 and 2.8 f1-scores respectively. Our work discusses the potential of counterfactual explanations to better estimate the accuracy of an AI model and reduce over-reliance on 'wrong' AI outputs and implications for improving human-AI collaborative decision-making.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; **User studies**; • **Applied computing** → **Health care information systems**; • **Computing methodologies** → *Artificial intelligence*; *Machine learning*.

Additional Key Words and Phrases: Human Centered AI; Human-AI Collaboration; Explainable AI; Trust; Reliance; Clinical Decision Support Systems; Physical Stroke Rehabilitation Assessment

1 INTRODUCTION

As advanced artificial intelligence (AI) and machine learning (ML) models have achieved equivalent results or outperformed humans at decision-making tasks (e.g. screening lung cancer [4]), these AI and ML models are increasingly being considered to increase efficiency and reduce the cost of performing decision-making tasks from various types of organizations and domains [59] (e.g. health [14, 42, 56], bail decisions [32], child welfare services [9], university admissions decisions [16], etc.). Specifically, researchers have discussed the potential of human and AI/ML teaming to achieve better results than either humans or AI/ML models alone [5, 42, 43]. However, previous research works have discussed that users might place too much trust in the AI/ML system and even agree with 'wrong' AI outputs [10, 11, 38].

Many researchers have discussed that the explainability [61] of a system is critical for human-AI collaborative decision-making [14, 42]. In particular, humans can review AI explanations to understand how the AI models generate outputs [61] and identify whether an AI output is right or not [14, 42]. There is growing number of studies to evaluate the effect of AI explanations on decision-making tasks [5, 10, 11, 14, 38, 42, 64]. For instance, researchers studied the effect of AI explanations on over-reliance using the simulated AI models or the tasks that do not require domain experts, such as judges or clinicians [5, 38, 64]. However, there has been contradictory perspectives

Authors' address: Min Hun Lee, mhlee@smu.edu.sg; Chong Jun Chew, cjchew.2020@scis.smu.edu.sg, Singapore Management University, Singapore, Singapore.

on the effect of AI explanations on user's trust on an AI output: users' trust in an algorithmic decision is not affected by the explanation interface [16] or can be increased by just presenting explanations [5].

In this work, we contribute to an empirical study that analyzes the effect of AI explanations on users' trust and reliance on AI during clinical decision-making. Specifically, we focus on the task of assessing post-stroke survivors' quality of motion [41]. Among various types of AI explanations [39, 61], this work explores salient feature analysis and counterfactual explanations for the following reasons. First, the previous research describes that therapists preferred to review feature-based explanations on rehabilitation assessment tasks [41]. However, the previous research discusses the issues of these feature-based explanations on overtrust in AI [29, 37, 64]. One potential reason for overreliance on AI might be that humans mostly employ heuristics and shortcuts while rarely involving analytical thinking during decision-making [10, 27]. Previous research [10] discusses the potential of cognitive forcing functions (e.g. not showing AI suggestions by default or waiting before showing AI suggestions) to increase analytical thinking and reduce overreliance. In this work, we assume that reviewing counterfactual explanations [47, 62, 64] will allow a user to critically think of how to change an AI output and more rigorously review an AI output than widely used AI explanations (e.g. feature-based or example-based explanations) that show information relevant to an AI output. We hypothesize that reviewing counterfactual explanations [47, 62, 64] will improve the user's analytical review of an AI output, assist the user to achieve better calibrated trust in AI, and reduce overreliance on it.

To this end, we conduct a within-subject experiment with seven therapists and ten laypersons to compare the effect of counterfactual explanations with one of the widely used AI explanations, salient feature explanations [5, 16, 38, 42]. Our results show that the human + AI team with both salient feature and counterfactual explanations improved the performance and agreement level on decision-making tasks only when 'right' AI outputs were presented. In contrast, when 'wrong' AI outputs were presented, the human + AI team with salient feature analysis had higher overreliance on 'wrong' AI outputs while the human + AI team with counterfactual explanations reduced overreliance on 'wrong' AI outputs by 21% compared to salient feature explanations.

When we analyzed the performance and the effect of AI explanations by therapists and laypersons, therapists had lower performance degradation and overreliance on 'wrong' AI outputs than laypersons: therapists' human + AI team performance was lower than their human alone performance by 8.6 f1-score with salient feature and 2.8 f1-score with counterfactual explanations; laypersons' human + AI team performance was lower than their human alone performance by 18.0 f1-score with salient feature and 14.0 f1-score with counterfactual explanations. Overall, reviewing counterfactual explanations assisted both therapists and laypersons to diversify their assessment (i.e. lower agreement level) and have more cases of rejecting 'wrong' AI outputs and fewer cases of agreeing with 'wrong' AI outputs than salient feature analysis by 19% from therapists and by 35% from laypersons.

When it comes to a self-reported usability score, therapists and laypersons had a higher self-reported trust score (73.78 out of 100) on the AI system with salient feature analysis than the AI system with counterfactual explanations (45.20 out of 100). The self-reported trust score of the system with counterfactual explanations is closer to the system performance (0.375: 3 right outputs of out 8) than that of the system with salient feature analysis.

Overall, this work provides new insights into the potential of counterfactual explanations to reduce overreliance on 'wrong' AI outputs and better estimate the performance of an AI model through a user study using uncontrolled AI outputs and explanations with therapists and laypersons on clinical decision-making tasks (i.e. rehabilitation assessment). In addition, our work compares the effect of AI outputs and explanations on domain experts and lay group participants. Our work

advances ongoing discussions around the implications for improving human-AI collaborative decision-making in high-stake domains (e.g. health) [15, 42, 44].

2 RELATED WORK

2.1 Towards Human-AI Collaborative Decision-Making

With the recent advance in artificial intelligence (AI) and machine learning algorithms, AI/ML models are increasingly being considered to assist humans' decision-making tasks in a variety of domains (e.g. health). Instead of applying fully autonomous AI systems, researchers have explored the feasibility of human-AI collaborative decision-making, in which an AI model provides humans new data-driven insights on a task for achieving complementary performance, outperforming neither of the AI or the human alone [5, 30, 42, 43]. For instance, a deep learning-based system has been used in clinics to bring new data-driven insights to assist the diagnosis of cancer [15], the detection of diabetic eye disease [8], or the assessment of physical stroke rehabilitation assessment [42].

Although previous research describes the potential of AI/ML systems to outperform human experts on prediction tasks [18, 32, 40], it still remains a challenge to develop and integrate these systems in practice due to the lack of human-centered designs and performing as a "black-box" system [13–15, 22, 31, 41, 65]. For the issue of lack of human-centered designs, there has been increasing recent research efforts [8, 15, 22, 41, 58, 65] that highlight the importance of involving stakeholders to understand their practices and needs [15, 41, 65] and socio-environmental factors [8] for the design and evaluation of a system. For instance, Yang et al. [65] conducted a field evaluation on the design of a decision support tool for cardiologists with synthetic data and found that clinicians are more likely to embrace a tool that augments their decision-making in natural and intuitive ways. Lee et al. [41] conducted interviews and focus-group sessions with therapists to understand the challenges and needs during rehabilitation assessment to design a human-centered decision support system.

2.2 User Studies of Explainable AI

In addition, researchers have discussed the importance of an AI explanation to communicate an AI output to a user [5, 20, 29] and realize human-AI collaboration [48] for a decision-making task [15, 42]. There has been a growing number of studies that have evaluated the effect of AI explanations in diverse decision-making tasks (e.g. house price prediction [55], image classification [2], student admission [16], deception detection [38], stroke rehabilitation assessment [42]) and aspects, such as whether an AI explanation assists a user to debug [29] or update an AI model [14, 42] or improves user's trust in AI [10, 14, 36, 52].

For instance, Alqaraawi et al. [2] conducted a user study on image classification tasks to evaluate the performance of the saliency map, an XAI technique that highlights input pixels in the original images that contribute to a model prediction and discussed the limited usefulness of saliency map to assist participants to anticipate a model output. Wang and Yin [64] conducted the randomized experiment using four types of common model-agnostic explainable AI methods and discussed the effect of AI explanations could be largely different where people have varying levels of domain expertise.

There have been contradictory perspectives on the effect of AI explanations on user's trust in an AI output: users' trust in an algorithmic decision is not affected by the explanation interface [16] or can be increased by just presenting explanations [5] or even when explanations are randomly chosen [38]. According to the study with MTruk worker on the task of deception detection task [38], Lai and Tan discussed that the presentation of AI-predicted labels and explanations improves human

performance on a task. In contrast, Bussone et al. discussed that providing richer explanations could lead to a harmful effect: overreliance on the system [11]. Along this issue, Buccinca et al. [10] discussed the cognitive forcing intervention, such as slowing down the process and asking the person to make a decision before seeing the AI recommendation, can reduce the overreliance on AI.

In addition to the contradictory perspectives on the effect of explanations on trust, our research community still requires additional studies to understand the effect of AI explanations on overreliance [54]. Specifically, even if Bussone et al. [11] and Buccinca et al. [10] investigated the effect of AI explanations on user's overreliance, previous research utilized a mock-up decision support system that operates with the wizard-of-oz approach [11] or a simulated AI model [10]. Other works that utilize AI/ML models focused on tasks that do not require domain experts, such as judges or clinicians [5, 38, 64].

In this work, we focused on the AI-assisted clinical decision-making task (i.e. physical stroke rehabilitation assessment) and investigate the effect of the salient feature and counterfactual explanations on users' trust and overreliance on AI. Specifically, this work utilized uncontrolled AI model outputs and explanations implemented by the dataset of 15 post-stroke survivors in contrast to existing previous research that utilizes simulated and controlled AI outputs and explanations [10, 11, 49] to understand the effect of AI explanations and the issue of overreliance. This work contributes to increasing knowledge on the effect of AI explanations by (i) comparing human alone and human + AI team in terms of performance, agreement level, and the number of 'right' or 'wrong' decisions and (ii) analyzing these evaluation metrics between domain experts (i.e. therapists) and laypersons. This work further discusses the potential of counterfactual explanations as a cognitive forcing function to better achieve a calibrated trust in AI and reduce overreliance on AI and implications for improving human-AI collaborative decision-making in high-stake domains (e.g. health) [15, 42, 44].

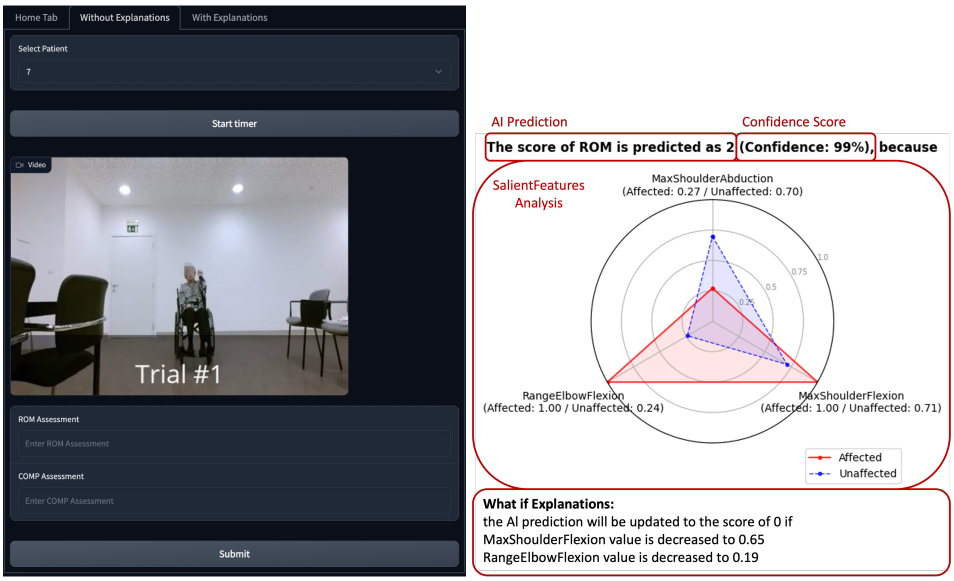
3 STUDY DESIGN

The primary research question of this work is to investigate the effect of AI explanations on users' trust and reliance on imperfect AI outputs. Building upon growing works on the usage of explainable AI methods for improving AI-assisted decision-making [10, 11, 64], we hypothesize that counterfactual explanations [47, 62], a type of AI explanations that describe how the inputs can be modified to achieve an AI output in a certain way, will increase user's analytic reviews and deliberations on an AI output and reduce user's overreliance on 'wrong' AI outputs. To this end, we conducted a within-subject experiment with therapists and laypersons in the context of assessing post-stroke survivors' quality of motion. Specifically, we compared the effect of using a decision support system with counterfactual explanations (Figure 1) to a baseline system with one of the widely used explainable AI techniques, salient features, calculating the importance of input features, [41, 45, 46]. Our study aims to explore the following research question:

- How do counterfactual explanations impact user's (1) performance & agreement level on decision-making tasks and (2) reliance and trust on AI outputs?

3.1 Clinical Decision Making Task: Physical Stroke Rehabilitation Assessment

In this work, we focus on a clinical decision-making task: assessing the quality of motion of patients affected by stroke, the second leading cause of death and third most common contributor to disability [19]. Building upon previous works on AI-assisted decision-making on physical stroke rehabilitation assessment [42], this work utilizes an upper-limb rehabilitation exercise (Figure 2) and two performance components of rehabilitation assessment: Range of Motion (ROM) and Compensation.



(a) (b)

Fig. 1. The AI-based decision support system that presents (a) the video of post-stroke survivor's exercises and (b) the predicted assessment score by AI along with salient feature-based explanations that compare a post-stroke survivor's unaffected and affected side using the top three most important features and counterfactual, what-if explanations that describe how input features need to be changed to flip an AI output (e.g. 'correct' to 'incorrect' ROM).

For an exercise, a post-stroke survivor has to raise his or her wrist to the mouth as if drinking water (Figure 2a). For the rehabilitation assessment, the *'ROM'* component refers to how closely a post-stroke survivor achieves the target position of an exercise (e.g. bring the wrist to the mouth) and the *'Compensation'* component indicates whether a post-stroke survivor involves any unnecessary joints to perform an exercise (e.g. leaning trunk to the side and backward - Figure 2b).

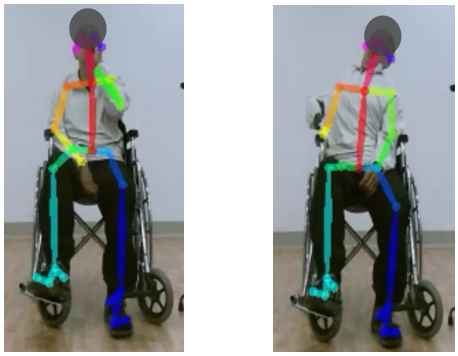
For the rehabilitation assessment task, participants went through the tutorial on rehabilitation assessment and were asked to review the video of post-stroke survivor's exercises and assess the post-stroke survivor's quality of motion in terms of the *'ROM'* and the *'Compensation'*. The score guidelines for rehabilitation assessment can be found in Table 3 in the Appendix.

3.2 Dataset, AI Models and Explanations

3.2.1 Dataset and Kinematic Features.

We utilized the dataset of a *'Bring a cup to the mouth'* upper-limb exercise from 15 post-stroke survivors [40]. Specifically, this dataset includes (1) the 300 videos of 15 post-stroke survivors performing the exercise (10 trials using their unaffected and affected side by stroke respectively), (2) their estimated joint positions using a Kinect sensor v2, and (3) the annotations by the expert therapist, who evaluated the post-stroke survivors' using clinically validated Fugl Meyer Assessment [60] and watched the recorded videos without reviewing any AI outputs.

Given the estimated joint positions of post-stroke survivors' exercises, we extracted various kinematic features based on the previous work [40, 60]. For the *'ROM'* component, we extracted joint angles (e.g. elbow flexion, shoulder flexion, elbow extension), normalized relative trajectory



(a) E1-Unaffected (b) E1-Affected

Fig. 2. Sample Unaffected and Affected Motions of an Exercise: (a) a patient can raise the patient's wrist to the mouth, (b) a patient compensated with trunk and shoulder joints.

(i.e. Euclidean distance between two joints - head and wrist, head and elbow), and normalized trajectory distance (i.e. the absolute distance between two joints - head and wrist, shoulder and wrist) in the x, y, z coordinates [40]. For the *'Compensation'* component, we extracted normalized trajectories (distances between joint positions of head, spine, and shoulder in the x, y, z coordinates from the initial to current frames) to distinguish the occurrence of a compensated movement [40, 60]

3.2.2 AI Models and Explanations.

We utilized a feed-forward Neural Network model to classify the quality of post-stroke survivor's motion due to its out-performance shown in the previous work [40]. Specifically, we grid-searched various architectures (i.e. one to three layers with 32, 64, 128, 256, 512 hidden units) and an adaptive learning rate with different initial learning rates (i.e. 0.0001, 0.005, 0.001, 0.01, 0.1) using cross-entropy loss and *'AdamOptimizer'* until the tolerance of optimization became 0.0001 or the maximum 200 iterations. We applied the leave-one-subject-out cross-validation, which trains a machine learning (ML) model with data from all post-stroke survivors except one post-stroke survivor, and test the model with the held-out post-stroke survivor. The model parameters that achieved the best F1 score during the cross-validation are described in the Appendix. Table 4. The trained ML model achieved an average F1-score of 0.9285 for the *'ROM'* and an average F1-score of 0.7867 for *'Compensation'* performance component.

After training ML models, we utilized widely used, open-source libraries to generate AI explanations: salient feature analysis and counterfactual explanations. Among various types of AI explanations, this work focuses on exploring salient feature analysis, building upon the previous research that described therapists' preferences in reviewing feature-based explanations on rehabilitation tasks [41]. However, the previous research describes the issues of these explanations on overtrust in AI [29, 37, 64]. This work assumes that counterfactual explanations will induce users to engage in a more critical review of an AI output by thinking about how to change an AI output compared to other widely used AI explanations (e.g. feature-based or example-based explanations) that provide relevant information on confirming an AI output. Thus, this work explores whether the counterfactual, what-if explanations [12] can assist users to better critically review AI outputs and explanations.

We utilized the SHAP [46, 57] for identifying salient features and the DiCE library [47] for generating the counterfactual explanations. For salient feature explanations (Figure 1b), we identified patient-specific, salient features and utilized only the top three salient features with the highest scores to avoid overwhelming users [35, 41]. For the presentation of these salient features, we utilized a radar chart to effectively show the comparison of identified features on post-stroke survivors' unaffected and affected sides to follow the therapist's practices [3, 41]. For instance, Figure 1b shows that the system identified *'MaxShoulderAbduction'*, *'RangeElbowFlexion'*, *'MaxShoulderFlexion'*, statistics of joint angles as the top three, most important features to assess the post-stroke survivor's quality of motion at Figure 1a. The radar chart describes the differences in identified feature values on post-stroke survivors' unaffected and affected sides.

The counterfactual explanations describe what changes in feature values lead to updating an AI output in a certain way [12, 24, 47]. To generate counterfactual explanations, we applied the model agnostic approach that utilizes the genetic algorithm [24, 47, 51] to find only three counterfactuals close to the query point. In addition, we specified the features to be changed in the DiCE library using the identified salient features by the SHAP library and their desired range using patients' held-out normal data to avoid generating varying and unfeasible explanations.

For the presentation of counterfactual explanations, as we already had a radar chart visualization to describe the comparison between unaffected and affected sides of a post-stroke survivor, we generated textual descriptions of the changes in feature values and AI outputs (Figure 1b). For instance, Figure 1b shows that the value of *'MaxShoulderAbduction'* and *'RangeElbowFlexion'* should be reduced to 0.65 and 0.19 respectively to update the output, predicted score of AI from 2 (i.e. full range of motion - ROM) to 0 (i.e. limited ROM).

3.3 Conditions & Task Specifications

3.3.1 Conditions.

In this work, we specified two conditions with two different AI explanations to understand their effects on the participants' decision-making task, the rehabilitation assessment of the post-stroke survivors.

- The first, baseline condition refers to an AI-based decision support system that presents videos of post-stroke survivor's exercises along with AI prediction scores and salient feature analysis (Figure 1b without what-if explanations).
- The second condition refers to the AI system that includes additional counterfactual explanations (Figure 1b) compared to the first condition.

In this work, we leverage feature-based explanations for therapists to find evidence and confirm their hypothetical assessment [63]. In addition, we explore counterfactual explanations for avoiding therapists' early confirmation [63] and inducing more analytical reviews on an AI output to reduce overreliance on AI. As previous work describes therapists preferred to review feature-based explanations to find evidence and confirm their assessment [41], we consider that features-based explanations are required by default for therapists to find evidence and confirm their assessment. In addition, counterfactual explanations are required as additional information that serves as a cognitive forcing function to reduce overreliance on AI. Thus, we included both salient feature analysis and counterfactual explanations in the second condition and compared the first and second conditions to understand the effect of counterfactual explanations on users' overreliance on AI.

In the study, we referred to interfaces as "Condition A" and "Condition B" to avoid biasing participants. We referred to these conditions respectively as the AI with salient features and the AI with counterfactual explanations for clarity throughout the paper. We implemented the web interface of each condition using the Gradio library [1] to conduct the user study. By default, our

web interface involves three strategies of cognitive forcing functions [10] on both Condition A and Condition B to reduce overreliance on AI. Specifically, we implemented the tab menus of 'Without Explanations' and 'With Explanations' (Figure 1), so that an AI output is not shown to the users from the beginning and allows a user to review AI outputs and explanations and update or confirm their assessment afterward [10]. In addition, our interface takes around a second to load an AI output and AI explanations instead of explicitly setting 30 seconds of waiting time [10]. Compared to Condition A, we included counterfactual explanations in Condition B and explore the effect of counterfactual explanations as a cognitive function.

3.3.2 Task Specifications.
To investigate the effect of AI explanations on users' overreliance on 'wrong' AI outputs, we utilized the trained ML models (Section 3.2.2) to select the cases of rehabilitation assessment for each condition. Specifically, we assigned cases with 3 'right' AI outputs and 5 'wrong' AI outputs on each condition.

3.4 Participants & Procedure

3.4.1 Participants.
Seven therapists (3 male and 4 female) with an average of 12.85 years of experience in stroke rehabilitation (Table 1). In addition, we recruited ten laypersons (7 male and 3 female; 2 graduate students and 8 undergraduate students) without experience in stroke rehabilitation to compare their performance [50] and reliance on AI with expert therapists. Participants were recruited through advertisements sent to hospitals, university staff & mailing lists, and the contacts of the research team.

Among the seven therapists, five of them are occupational therapists whose primary roles are to help patients better engage in their daily activities. The two remaining therapists are physiotherapists who treat their patient's physical impairments from a bio-mechanical perspective. The detailed demographic information of participants is described in the Appendix (Table 5).

Table 1. Demographics of participants (therapists and laypersons).

ID	Role	Years in the Role	Q. Tech Experience	ID	Q. Tech Experience
T1	Occupational Therapist (OT)	25	2.6 +/- 2.0	L1	6.6 +/- 0.5
T2	Occupational Therapist (OT)	5	5.4 +/- 2.2	L2	5.8 +/- 0.7
T3	Occupational Therapist (OT)	10	4.0 +/- 2.4	L3	5.0 +/- 1.7
T4	Occupational Therapist (OT)	6	3.6 +/- 2.8	L4	5.6 +/- 1.5
T5	PhysioTherapist (PT)	17	3.6 +/- 2.3	L5	3.0 +/- 2.1
T6	Occupational Therapist (OT)	12	5.2 +/- 2.1	L6	6.0 +/- 0.0
T7	PhysioTherapist (PT)	15	3.2 +/- 1.0	L7	4.6 +/- 2.6
				L8	5.0 +/- 1.3
				L9	5.2 +/- 1.9
				L10	5.2 +/- 1.8

To understand the participant's background in technology, we asked them to respond to a set of technical experience questions, which were based on survey questions designed by the Center for Research and Education on Aging and Technology Enhancement (CREATE) [17]. Each participant rated his or her experience with diverse recent technologies (i.e. computer/laptop, activity tracker, virtual voice assistant, unmanned convenient store, autonomous vehicle) on a 7-point scale (1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = neutral, 5 = somewhat agree, 6 = agree, 7

= strongly agree. A low score on technology experience (e.g. 1.0) indicates that a participant barely has experience with recent technologies. Overall, therapists have diverse levels of experience with recent technologies with an average score of 3.94 out of 7.0 and laypersons have a slightly higher average score of 5.2 out of 7.0.

3.4.2 Procedure.
The study was conducted online. After a participant completed the informed consent form that was approved by the Institutional Review Board, the participant went through the tutorial on rehabilitation assessment and the study procedure. Each participant was randomly assigned to either first use the AI with salient feature analysis **(Condition A - Features)** and then AI with salient features and counterfactual explanations **(Condition B - Countfacts)** or vice-versa. Each condition involves two sub-tasks. Specifically, we asked the participant to (a) first provide their initial assessment (Figure 1a) without AI outputs and explanations and (b) then finalized the assessment after reviewing AI outputs and explanations to understand the effect of reviewing AI outputs and explanations (Figure 3). In each condition, a participant was required to perform 8 decision-makings on rehabilitation assessment after reviewing post-stroke survivor's exercises. The sub-tasks of each condition were counterbalanced and the order of the two conditions and the presentations of post-stroke videos were randomized. After completing assessment tasks on each condition, the participant responded to the usability questions. After finishing all tasks on two conditions, the participant filled out the overall preference questionnaire. All participants received a fixed compensation for their participation in the study.

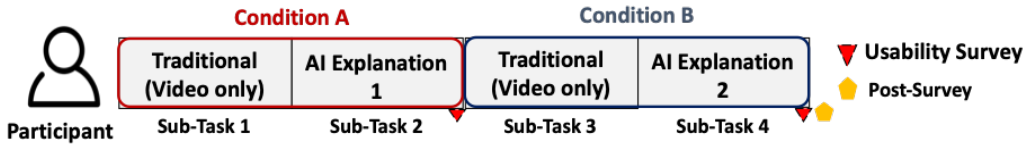


Fig. 3. The overall procedure of the user study: a participant completed rehabilitation assessment tasks using the AI with salient features (Condition A) and the AI with salient features and counterfactual explanations (Condition B). In each condition, the participant first completed the initial assessment after reviewing only a video and then provided the final assessment after reviewing the AI outputs and explanations. When the participant completed the tasks on each condition, the participant completed the usability questionnaires on each condition. At the end of the study, the participant completed the overall, post-survey about their preferences.

3.5 Data Analysis Metrics

We analyzed two systems (i.e. AI with salient feature analysis and AI with counterfactual explanations) using the following metrics: 1) performance and 2) participants' agreement level on rehabilitation assessment tasks, 3) counts of 'right' and 'wrong' decisions (including overreliance), 4) the duration of decision-making tasks, and 5) usability questionnaires [11, 37, 42].

3.5.1 Performance.
One of the most commonly used metrics on human-AI collaborative decision-making tasks is performance, measuring the percentage of correctly making decisions on instances [37]. In this study, we utilized the annotations of a therapist from the dataset [40] as ground truths and evaluate participants' performance on decision-making tasks before/after reviewing AI outputs and explanations.

3.5.2 Agreement Level.
Most medical diagnoses rely on standardized guidelines [23, 26, 60]. However, clinicians can be biased in their decision making and expert disagreement is prevalent in medical decision-making tasks [7, 33, 34]. Thus, we also analyzed the agreement level of participants' decisions before/after reviewing AI outputs and explanations.

3.5.3 Counts of 'right' and 'wrong' decisions.
In addition to the performance and agreement level, we analyzed the counts of 'right' and 'wrong' decisions by participants to further analyze their overreliance on 'wrong' AI outputs. Also, we measured the count of (1) agreeing with 'right' AI outputs, (2) rejecting 'wrong' AI outputs, (3) agreeing with 'wrong' AI outputs (i.e. overreliance), and (4) rejecting 'right' AI outputs for further analysis. In addition, we analyzed the number of times when AI explanations assisted participants to change and make 'right' or 'wrong' decisions.

3.5.4 Duration of Decision Making.
Our web interface measured the estimated duration of each decision-making by asking the participants to indicate their starting point of a decision-making task on the interface.

3.5.5 Usability Questionnaires.
We also utilized participants' self-reported, subjective responses on usability aspects of the systems with salient feature analysis and counterfactual explanations, building upon previous research of human-AI collaborative decision-making in health [14, 37, 41]. Specifically, these usability aspects include (1) Useful, (2) Insight, (3) Effort, (4) Transparent, (5) Trust, (6) Frustration, (7) UsageIntent, (8) AIPotential and (9) Preference between two interfaces as follows:

- Useful: *"The system provided useful information to understand patient's performance for assessment"* [14, 41].
- Insight: *"The system provided new insights on patient's performance for assessment"* [41].
- LessEffort: *"The system helped me think through and complete the assessment tasks with less effort"* based on the effort dimension of the NASA-TLX [25]
- Reliance: *"I relied on assessment scores & analysis from the system for my final assessment"*
- Transparent: *"The system was transparent about why it provided a particular assessment score"*
- Trust: *"I can trust the provided assessment scores or/and analysis from the system"*
- Frustration: *"I was insecure, discouraged, and stressed while using the system"* based on the frustration dimension of the NASA-TLX [25]
- UsageIntent: *"I would use this system to understand and assess patient's exercise performance in practice"* [14, 41].
- AIPotential: *"I think AI, data-driven tool can improve rehabilitation assessment"*
- Preference between two interfaces: participants rated on a 7-point scale ranging from 1 (totally Condition A), 2 (much more Condition A than B), 3 (slightly more Condition A than B), 4 (neutral), ..., 7 (totally Condition B) [14, 41].

All questionnaires were rated on a 7-point scale except for the trust, which was rated on a 100-point scale.

4 RESULTS

Throughout this paper, we refer to the outcomes of participants, who reviewed the videos without AI outputs and explanations as ***"Human"*** and those , who reviewed the videos with AI outputs and explanations as ***"Human + AI"***. Also, we refer the Condition A as ***"Features"***, in which participants

use the AI with salient feature analysis and the Condition B as ***"Counterfacts"***, where participants use the AI with salient features and counterfactual explanations.

For the performance and agreement level metrics, we analyzed the differences in outcomes between ***"Human"*** and ***"Human + AI"*** over two conditions. For the counts of *'right'* and *'wrong'* decisions, duration of decision makings, and usability questionnaires, we compared the outcomes of two conditions using AI with salient feature analysis and counterfactual explanations respectively.

In the following section, we reported the descriptive statistic of each metric and conducted the Wilcox significant tests using data from therapists and laypersons respectively. If outcomes from therapists and laypersons have the same trends, we also described the overall outcomes of data analysis metrics.

4.1 Performance

Figure 4 summarizes the average performance (i.e. F1-score) of rehabilitation assessment tasks by therapists (TPs) and laypersons (LPs) respectively.

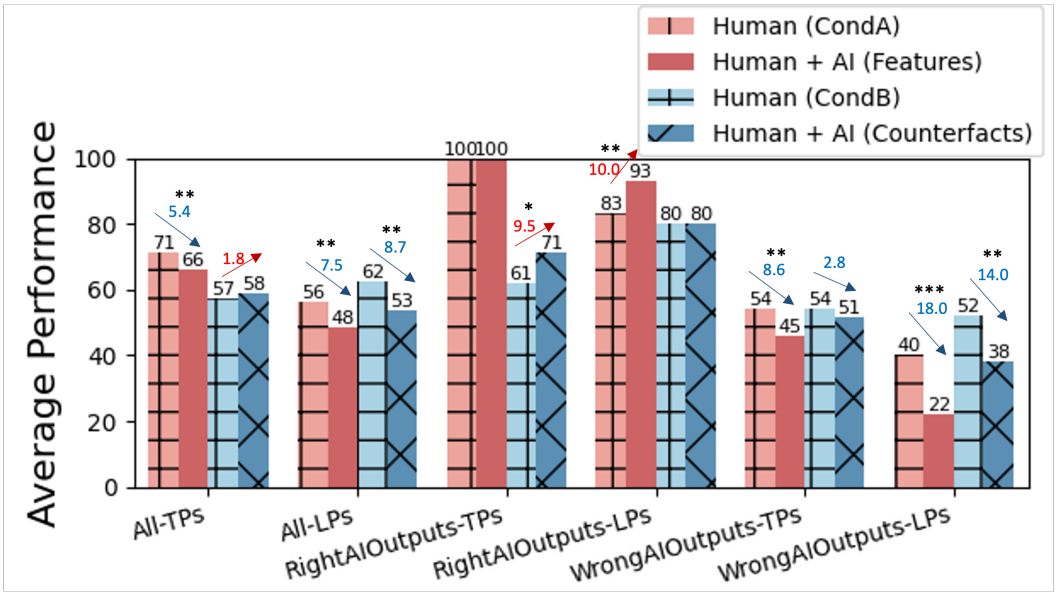


Fig. 4. Performance of rehabilitation assessment tasks by therapists (TPs) and laypersons (LPs) using AI with Features and Counterfactuals for (1) all cases, (2) cases with 'right' AI outputs, and (3) cases with 'wrong' AI outputs. Although participants' performances improved after reviewing 'right' AI outputs with both salient feature analysis and counterfactual explanations, their performance reduced after reviewing 'wrong' AI outputs. Counterfactual explanations assisted participants to have lower degraded performance than salient feature analysis. *, **, and *** indicate 90%, 95%, and 99% statistical significance levels.

Overall, therapists' and laypersons' human + AI team performance with both salient feature analysis and counterfactual explanations were lower than their human alone performance ($p < 0.05$) except for a marginal improvement of therapists' human + AI team performance. For further analysis, we analyzed the performances of therapists and laypersons using the cases with 'right' or 'wrong' AI outputs. When 'right' AI outputs were presented to therapists and laypersons, therapists' human + AI team performance with counterfactual explanations and laypersons' human + AI team performance with salient features were higher than their human alone performance ($p < 0.1$ and

$p < 0.05$ respectively). However, when 'wrong' AI outputs were presented to the therapists and laypersons, their human + AI team performance with salient features or counterfactual explanations was decreased. Compared to the therapists' performances, laypersons' performances were degraded significantly. Also, we found that therapists' and laypersons' human + AI team performances with salient features (i.e. 8.6 and 18.0 F1-scores respectively) led to higher performance degradation than their performance with counterfactual explanations (i.e. 2.8 and 14.0 F1-scores respectively).

4.2 Agreement Level

Figure 5 summarizes an average agreement level (e.g. F1-score) of rehabilitation assessment tasks by therapists (TPs) and laypersons (LPs) respectively.

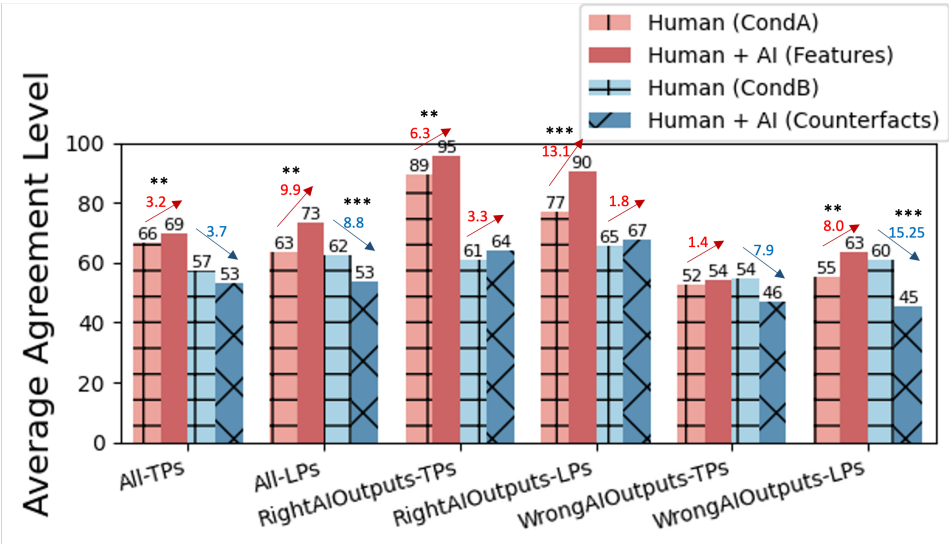


Fig. 5. Agreement level of rehabilitation assessment tasks by therapists (TPs) and laypersons (LPs) using AI with Features and Counterfactuals for (1) all cases, (2) cases with 'right' AI outputs, and (3) cases with 'wrong' AI outputs. After reviewing 'right' AI outputs with salient feature analysis or counterfactual explanations, both TPs and LPs increased their agreement levels. After reviewing 'wrong' AI outputs, both TPs and LPs using salient features increased their agreement levels while TPs and LPs using counterfactual explanations decreased their agreement levels. *, **, and *** indicate 90%, 95%, and 99% statistical significance levels.

Overall, both TPs and LPs increased their agreement level when they reviewed AI outputs (Human + AI) of salient features ($p < 0.05$) and decreased their agreement level when they reviewed AI outputs of counterfactual explanations.

For the cases with 'right' AI outputs, both TPs and LPs achieved higher agreement levels with statistical significance ($p < 0.05$ and $p < 0.01$ respectively) when they used salient features. Also, they achieved higher agreement levels without significance when they used counterfactual explanations. For the cases with 'wrong' AI outputs, they increased their agreement levels (i.e. by 1.4 F1-score for TPs; by 8.0 F1-score for LPs) when they reviewed salient features. In contrast, they decreased their agreement levels when they reviewed counterfactual explanations. Similar to the performance metrics, an agreement level of TPs using counterfactual explanations (counterfacts) led to lower degradation of the agreement level (i.e. -7.9% F1-score) than that of LPs using counterfacts (i.e. a −15.25% F1-score, $p < 0.01$).

4.3 Counts of 'Right' and 'Wrong' Decisions

Figure 6 summarizes the counts of participants' 'right' and 'wrong' decisions. Overall, the human + AI team with counterfactual explanations by therapists (TPs) and laypersons (LPs) had more cases of 'right' decisions than the human + AI team with salient feature analysis: 21% (29 out 136) from all participants ($p < 0.01$), 8% (5 out of 56) from TPs ($p < 0.1$), and 30% (24 out of 80) from LPs ($p < 0.01$). In addition, the human + AI team with counterfactual explanations had fewer cases of 'wrong' decisions than the human + AI team with salient feature analysis: 21% (29 out 136) from all participants ($p < 0.01$), 8% (5 out of 56) from TPs ($p < 0.1$), and 30% (24 out of 80) from LPs ($p < 0.01$).

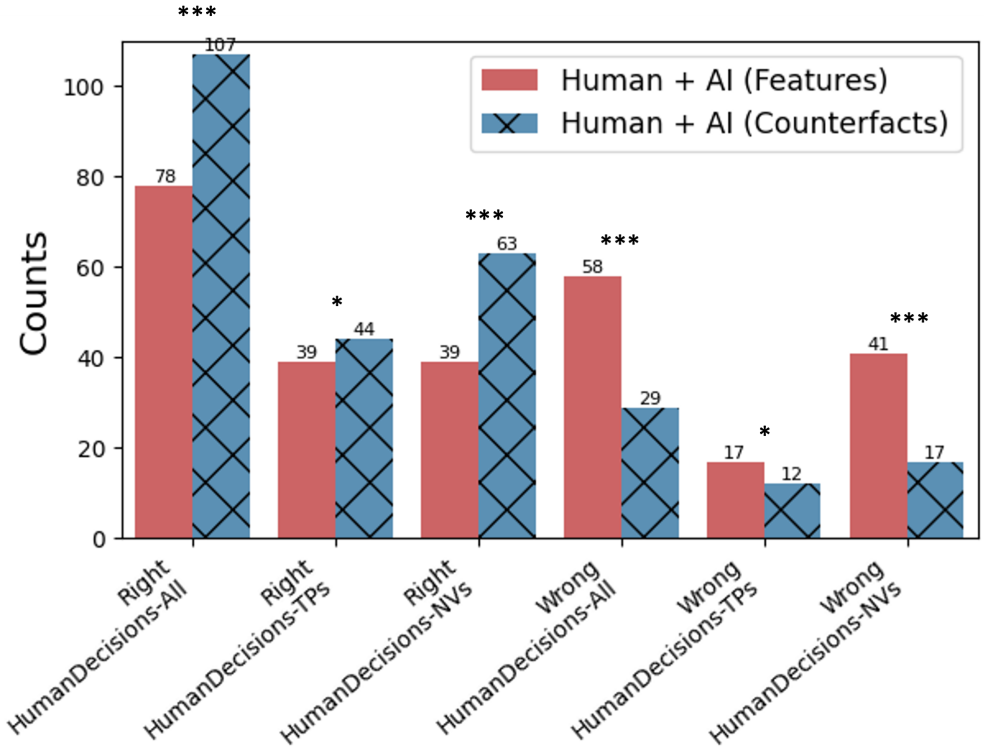
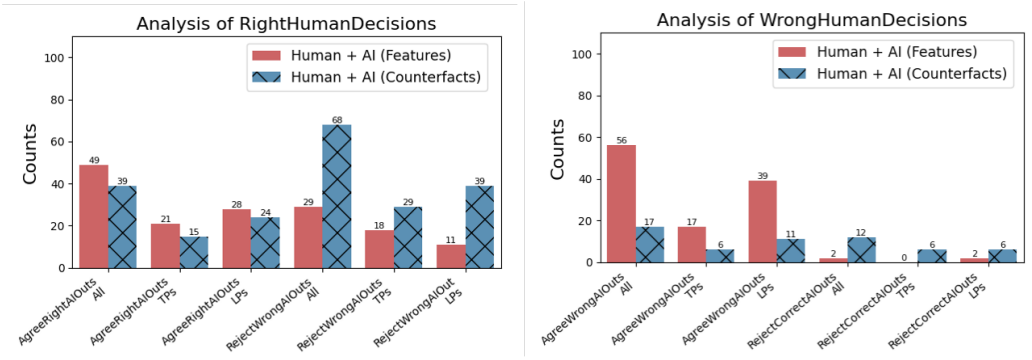


Fig. 6. The counts of 'right' and 'wrong' decisions by all participants (All), therapists (TPs), and laypersons (LPs). Counterfactual explanations assisted the participants to increase their 'right' decisions and reduce their 'wrong' decisions compared to the salient feature analysis. * and *** indicates 90% and 99% statistical significance levels.

For the detailed analysis, we analyzed the number of 'right' and 'wrong' decisions of TPs and LPs by (1) agreeing with 'right' AI outputs, (2) rejecting 'wrong' AI outputs, (3) agreeing with 'wrong' AI outputs, and (4) rejecting 'right' AI outputs (Figure 7). The human + AI team with counterfactual explanations had more cases of rejecting 'wrong' AI outputs and fewer cases of agreeing with 'wrong' AI outputs than the human + AI team with salient features: by 19% (11 out of 56) from TPs and by 35% (28 out of 80) from LPs. In addition, the human + AI team with Counterfacts had fewer cases of agreeing with 'right' AI outputs and more cases of rejecting 'right' AI outputs than the human + AI team with Features: by 10% (6 out of 56) from TPs and by 5% (4 out of 80) from LPs.



(a) (b)

Fig. 7. Detailed analysis of (a) 'right' and (b) 'wrong' human decisions by all participants (All), therapists (TPs), and laypersons (LPs): Counterfactual explanations assisted the participants to increase the number of 'right' decisions on rejecting 'wrong' AI outputs and reduce the number of 'wrong' decisions on agreeing 'wrong' AI outputs.

4.4 Duration of Decision Making

The participants took an average of 57 seconds (All), 49 seconds (TPs), and 63 seconds (LPs) using the system with salient feature analysis and an average of 75 seconds (All), 70 seconds (TPs), and 80 seconds (LPs) using the system with counterfactual explanations to complete a single decision-making task. Overall, the system with counterfactual explanations requires an average of 18 more seconds (All), 21 more seconds (TPs), and 17 more seconds (LPs) than the system with salient feature analysis on a decision-making task.

4.5 Usability Questionnaires

Figure 8 summarizes the usability responses by the participants using the system with (1) salient feature analysis (Features) and (2) counterfactual explanations (Counterfacts).

The participants (both therapists and laypersons) considered that the system with Features is more useful ($p < 0.01$), provides more insights without statistical significance, requires less effort ($p < 0.01$), more reliable ($p < 0.01$), more transparent without statistical significance, more trustful ($p < 0.01$), less frustrating ($p < 0.01$). Overall, they both expressed higher usage intent ($p < 0.01$) and higher potential ($p < 0.01$) of the system with Features than the system with Counterfacts.

For the post-survey on the preference question, Table 2 describes that there are 7 participants, who preferred the system with salient feature analysis (6 totally; 1 much more; 1 slightly more), 7 participants, who preferred the system with counterfactual explanations (2 totally; 3 much more; 3 slightly), and 1 neutral.

Participants preferred to use the system with salient feature analysis as its visualization is *"faster to read and process information"* (TP1) than textual, counterfactual explanations even if the other system. Other participants preferred the system with counterfactual explanations because they considered that these explanations assisted to *"provide a second view to help assessment"* (TP 2) and *"confirm any doubt during the assessment"* (TP 6).

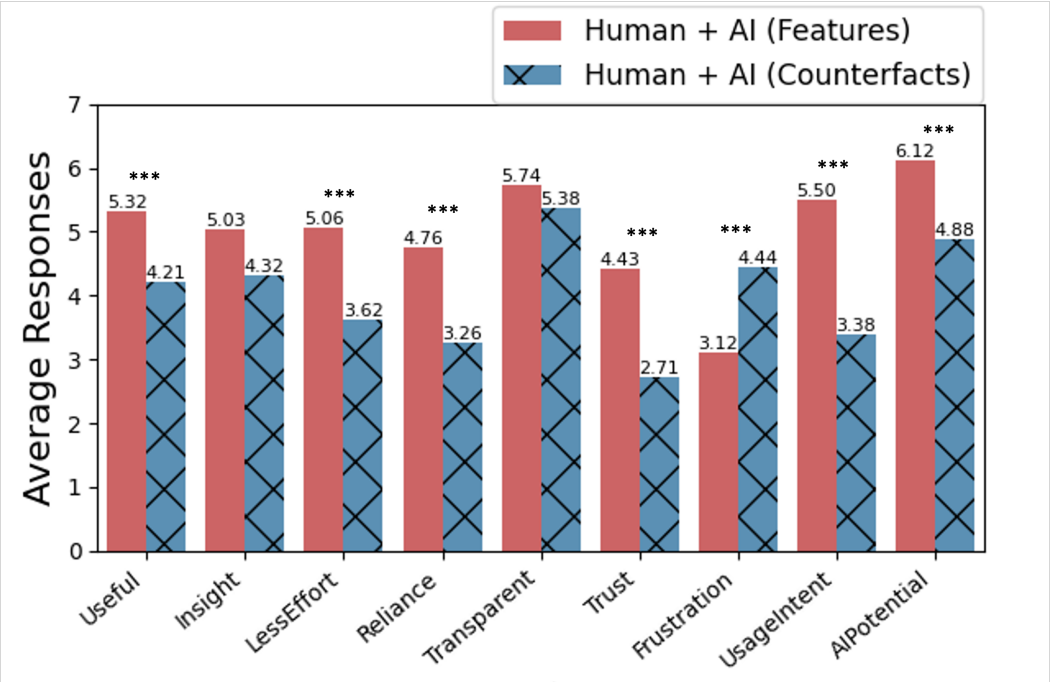


Fig. 8. Participants' usability responses on the system with (1) salient feature analysis (Features) and (2) counterfactual explanations (Counterfacts). Overall, participants expressed higher usage intent and potential with the system with salient feature analysis. They considered that the system with salient feature analysis is considered to be more useful, provides more insights without significance, requires less effort on assessment, is more reliable, transparent without significance, more trustful, and less frustrating. *** indicates 99% statistical significance level.

Table 2. Participants' preferences on the system: overall, there are 7 participants, who preferred version A, salient feature analysis (6 totally, 1 much more, 1 slightly more), 7 participants, who preferred version B, counterfactual explanations (2 totally, 3 much more, 3 slightly more), and 1 neutral.

UserType	(1) Totally version A	(2) Much more version A than B	(3) Slightly more version A than B	(4) Neutral	(5) Slightly more version B than A	(6) Much more version B than A	(7) Totally version B
Therapists	2	1	1	0	1	2	1
Laypersons	4	0	0	1	2	1	1
Overall	6	1	1	1	3	3	2

5 DISCUSSION

In this section, we discussed the potential benefits and limitations of two types of AI explanations (i.e. salient feature and counterfactual), their effects on domain experts and laypersons' decision-making, and suggestions for more effect human-AI collaborative, clinical decision-making.

5.1 Effects of Salient Feature & Counterfactual Explanations for Overreliance on AI

AI explanations have been considered as an important communication medium to realize effective human-AI collaborative decision-making tasks [37, 42]. In contrast to prior research that describes

the improved performance of humans with AI explanations [38, 41], our results demonstrated that the human + AI team with both salient feature and counterfactual explanations performed worse than the human alone (Figure 4).

According to the further analysis of the cases with 'right' and 'wrong' AI outputs, the presentation of AI outputs and explanations has different effects on the human + AI team performance (Figure 4). Specifically, when 'right' AI outputs are presented, the human + AI team with both salient feature and counterfactual explanations performed better than the human alone. However, the human + AI team with salient feature explanations and counterfactual explanations performed worse than the human alone.

Compared to the human + AI team with salient feature explanations, the human + AI team with counterfactual explanations supported both therapists and laypersons to have more 'right' decisions (21%: 29 out of 136) and fewer 'wrong' decisions (21%: 29 out of 136) (Figure 6). Overall, our findings indicate user's overreliance on 'wrong' AI outputs, which follows the previous studies that describe salient feature explanations increases user's overreliance on the AI model [6, 64]. In addition, our results show that counterfactual explanations performed better than salient feature explanations to assist reduce therapists' and laypersons' overreliance on AI.

When it comes to the agreement level, our results showed that the human + AI team with salient feature analysis led to an increase in the agreement level on the cases with 'right' and 'wrong' AI outputs (Figure 5). However, we found that the increase in agreement level does not necessarily indicate a positive performance improvement. Specifically, the counts of participants' 'right' and 'wrong' decisions (Figures 6 and 7) indicated that the human + AI team with salient feature analysis had more 'wrong' decisions while having a lower number of rejecting 'wrong' AI outputs and a higher number of agreeing 'wrong' AI outputs than the human AI team with counterfactual explanations. Taken together, our findings show that counterfactual explanations can serve as a cognitive forcing function [10] that assists the users in analytically reviewing AI explanations and reducing their overreliance on 'wrong' AI outputs.

5.2 Domain Experts vs Laypersons

Among various data analysis metrics, we found that both therapists and laypersons had mostly the same outcome patterns in performance, agreement level, counts of 'right' and 'wrong' decisions, duration of decision-making, and usability responses. However, our results suggest that laypersons had a higher over-reliance on AI outputs than therapists.

Specifically, when 'wrong' AI outputs were presented, laypersons had much higher performance degradation by 18.0 f1-score with salient feature explanations and 14.0 f1-score with counterfactual explanations than therapists who had a performance degradation of 8.6 f1-score with salient feature explanations and 2.8 f1-score with counterfactual explanations. In addition, this over-reliance on 'wrong' AI outputs has shown more significance with laypersons than domain experts, therapists.

5.3 Towards Better Calibrated Trust and Evaluation Metrics on AI

Similar to the previous research [11], our study also shows a positive correlation between user's trust and reliance: the more a user trusts the system, the more the user is likely to over-rely on outputs of the system even when 'wrong' AI outputs are presented. Although we provided the same number of 'right' and 'wrong' AI outputs on two systems with salient feature analysis and counterfactual explanations, participants considered that the system with salient feature analysis "*is more accurate*" (TP 6) than the system with counterfactual explanations.

Overall, participants expressed that they had a higher, self-reported trust score and a higher reliance score on the system with salient feature analysis than the system with counterfactual

explanations. In particular, the trust score of the system with salient feature analysis is 73.76 out of 100 and that of the system with counterfactual explanation is 45.20 out of 100.

As our task specification (Section 3.3.2) includes 3 'right' and 5 'wrong' AI outputs, the ideal estimation of an ML performance is 0.375 (3 out of 8). The participants using the system with neither explanation exactly estimated the performance of an ML model. However, the trust score of the system with counterfactual explanation is much closer than that of the system with salient feature analysis. Thus, this finding suggests that counterfactual explanations also have the potential to assist a user in better evaluating and estimating the accuracy of an ML model.

In addition, our findings suggest that possible gaps between users' perceived benefits and actual trustworthiness of an AI system. Relying only on subjective usability responses [37] might be limited and does not provide an appropriate understanding and evaluation on the trustworthiness of an AI system. In other words, an AI system with a higher self-reported trust score by participants does not necessarily mean that the system would achieve human + AI complementary team performance. It is important to explore a way or metrics to more accurately evaluate the trustworthiness and effectiveness of AI systems in the future.

5.4 Limitations

Our results demonstrated the potential of the human + AI team with counterfactual explanations to reduce the overreliance on AI and make the users better estimate the accuracy of an AI model during human-AI collaborative decision-making using uncontrolled AI outputs and explanations that are more stochastic. However, participants had lower usage intent and expected lower potential of the AI system with counterfactual explanations as reviewing a counterfactual explanation presented in texts *"could be more confusing"* (T1) and *"take more effort to complete the assessment"* (T4).

As previous research shows the higher understandability of counterfactual explanations by clinicians by including visual graphics than textual descriptions [49], we believe our limited scores of usability aspects on counterfactual explanations might be overcome by exploring new visualizations and human-centered design of AI explanations [21, 28].

As this work primarily focuses on exploring the effect of counterfactual explanations on user trust, the experimental designs of this work do not consider the possible effect of explanation fidelity. It is important to further investigate the effect of explanation fidelity on user trust [53] and overreliance. In addition, an additional study is required to investigate how people can effectively evaluate the performance and trustworthiness of an ML model and calibrate their trust and reliance to improve human-AI/algorithm interaction [22].

Our work also has a limitation in its generalizability as our work does not involve a large number of participants. However, such a small sample size is not unusual in similar previous works [11, 41]. In addition, this study mainly explores our research question in the context of a single clinical decision-making task (i.e. rehabilitation assessment) and is limited by particular types and visualization formats of AI explanations and an ML model (i.e. a feed-forward neural network). It is required to conduct additional studies to explore other decision-making tasks and types of ML models, explanations, and visualizations [21, 28] for further generalization of our findings.

6 CONCLUSION

In this work, we contributed to an empirical study of analyzing the effect of the salient feature and counterfactual explanations on users' trust and reliance on AI during a human-AI collaborative clinical decision-making task (i.e. assessing post-stroke survivor's quality of motion). Our results showed that the humans + AI team with both salient feature and counterfactual explanations increased its performance on decision-making tasks only when 'right' AI outputs are presented and decreased its performance when 'wrong' AI outputs are presented. Our results demonstrated

that counterfactual explanations assisted the participants to reduce their overreliance on 'wrong' AI outputs (21 %) compared to salient feature explanations. Also, we found that laypersons had higher performance degradation and overreliance than domain experts, therapists. Taken together, our work brings to light that providing AI explanations does not necessarily indicate improved human-AI collaborative decision-making. This work discusses the potential of counterfactual explanations to improve analytical reviews on AI outputs to better estimate AI performance and reduce overreliance on AI with the cost of cognitive burdens and other implications for improving human-AI collaborative decision-making.

ACKNOWLEDGMENTS

The authors thank all the participants in this work for their time and valuable inputs. We also thank the anonymous reviewers for their constructive feedback. This work is supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

REFERENCES

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569* (2019).

[2] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[3] 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*. ACM.

[4] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine* 25, 6 (2019), 954–961.

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[6] 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.

[7] Michael L Barnett, Dhruv Boddupalli, Shantanu Nundy, and David W Bates. 2019. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA network open* 2, 3 (2019).

[8] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[9] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[10] 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.

[11] 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.

[12] Ruth MJ Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning.. In *IJCAI*. 6276–6282.

[13] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. 2017. Unintended consequences of machine learning in medicine. *Jama* 318, 6 (2017), 517–518.

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

[15] 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.

[16] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[17] Sara J Czaja, Neil Charness, Arthur D Fisk, Christopher Hertzog, Sankaran N Nair, Wendy A Rogers, and Joseph Sharit. 2006. Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (CREATE). *Psychology and aging* 21, 2 (2006), 333.

[18] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

[19] Valery L Feigin, Bo Norrving, and George A Mensah. 2017. Global burden of stroke. *Circulation research* 120, 3 (2017), 439–448.

[20] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.

[21] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 531–535.

[22] 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.

[23] Early Treatment Diabetic Retinopathy Study Research Group et al. 1987. Treatment techniques and clinical guidelines for photocoagulation of diabetic macular edema: Early Treatment Diabetic Retinopathy Study report number 2. *Ophthalmology* 94, 7 (1987), 761–774.

[24] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23.

[25] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[26] Joseph Jankovic. 2008. Parkinson's disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry* 79, 4 (2008), 368–376.

[27] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

[28] Smiti Kaul, David Borland, Nan Cao, and David Gotz. 2021. Improving Visualization Interpretation Using Counterfactuals. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 998–1008.

[29] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[30] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving human-AI partnerships in child welfare: understanding worker practices, challenges, and desires for algorithmic decision support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[31] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics* 6, 2 (2018), e24.

[32] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[33] Peter Knapp and Jenny Hewison. 1999. Disagreement in patient and carer assessment of functional abilities after stroke. *Stroke* 30, 5 (1999), 934–938.

[34] Jonathan Krause, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S Corrado, Lily Peng, and Dale R Webster. 2018. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology* 125, 8 (2018), 1264–1272.

[35] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[36] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[37] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[38] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*.

29–38.

[39] Min Hun Lee and Yi Jing Choy. 2023. Exploring a Gradient-based Explainable AI Technique for Time-Series Data: A Case Study of Assessing Stroke Rehabilitation Exercises. In *ICLR 2023 Workshop on Time Series Representation Learning for Health*.

[40] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2019. Learning to assess the quality of stroke rehabilitation exercises. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 218–228.

[41] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

[42] 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.

[43] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2022. Towards efficient annotations for a human-ai collaborative, clinical decision support system: A case study on physical stroke rehabilitation assessment. In *27th International Conference on Intelligent User Interfaces*. 4–14.

[44] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

[45] 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.

[46] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[47] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.

[48] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876* (2019).

[49] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2023. How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies* 169 (2023), 102941.

[50] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[51] J Arturo Olvera-López, J Ariel Carrasco-Ochoa, J Martínez-Trinidad, and Josef Kittler. 2010. A review of instance selection methods. *Artificial Intelligence Review* 34, 2 (2010), 133–143.

[52] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *CHI Conference on Human Factors in Computing Systems*. 1–9.

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

[54] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[55] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[56] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature Medicine* 28, 1 (2022), 31–38.

[57] Khushnaseeb Roshan and Aasim Zafar. 2021. Utilizing XAI technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (SHAP). *arXiv preprint arXiv:2112.08442* (2021).

[58] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. " The human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 99–109.

[59] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, et al. 2016. Artificial intelligence and life in 2030: the one hundred year study on artificial intelligence. (2016).

[60] Katherine J Sullivan, Julie K Tilson, Steven Y Cen, Dorian K Rose, Julie Hershberg, Anita Correa, Joann Gallichio, Molly McLeod, Craig Moore, Samuel S Wu, et al. 2011. Fugl-Meyer assessment of sensorimotor function after stroke: standardized training procedure for clinical practice and clinical trials. *Stroke* 42, 2 (2011), 427–432.

[61] Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems* 32, 11 (2020), 4793–4813.

[62] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

[63] 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.

[64] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[65] 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*. ACM, 238.

Table 3. Guidelines to Assess Stroke Rehabilitation Exercises

Performance Components	**Score**	**Guidelines**
Range of Movement (ROM)	0	Does not or barely involve any movement
	1	Less than half way aligned with an *'Target'* position
	2	Movement achieves an *'Target'* position
Compensation	0	Noticeable compensation in more than two joints
	1	Noticeable compensation in a joint
	2	Does not involve any compensations

Table 4. Parameters of Machine Learning Models (i.e. Feed-Forward Neural Network Models)

	Hidden Layers and Units / Learning Rate	
	ROM	Comp
E1	(256) / 0.005	(16, 16) / 0.01

Table 5. Detailed Demographics of Therapists (T1 - T7) and Laypersons (L1 - L10)

ID	Gender	Age	Q. Tech Experience	Occupation	Years in the Role
T1	Female	45 - 54 years	2.6 +/- 2.0	Occupational Therapist (OT)	25
T2	Female	25 - 34 years	5.4 +/- 2.2	Occupational Therapist (OT)	5
T3	Female	25 - 34 years	4.0 +/- 2.4	Occupational Therapist (OT)	10
T4	Male	25 - 34 years	3.6 +/- 2.8	Occupational Therapist (OT)	6
T5	Male	35 - 44 years	3.6 +/- 2.3	PhysioTherapist (PT)	17
T6	Female	25 - 34 years	5.2 +/- 2.1	Occupational Therapist (OT)	12
T7	Male	35 - 44 years	3.2 +/- 1.0	PhysioTherapist (PT)	15
L1	Female	25 - 34 years	6.6 +/- 0.5	Graduate Student	
L2	Female	25 - 34 years	5.8 +/- 0.7	Graduate Student	
L3	Male	18 - 24 years	5.0 +/- 1.7	Undergraduate Student	
L4	Male	18 - 24 years	5.6 +/- 1.5	Undergraduate Student	
L5	Male	18 - 24 years	3.0 +/- 2.1	Undergraduate Student	
L6	Male	18 - 24 years	6.0 +/- 0.0	Undergraduate Student	
L7	Male	18 - 24 years	4.6 +/- 2.6	Undergraduate Student	
L8	Male	18 - 24 years	5.0 +/- 1.3	Undergraduate Student	
L9	Male	18 - 24 years	5.2 +/- 1.9	Undergraduate Student	
L10	Female	18 - 24 years	5.2 +/- 1.8	Undergraduate Student	

Received January 2023; revised April 2023; accepted May 2023