A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment

Min Hun Lee
Carnegie Mellon University
minhunl@cs.cmu.edu

Daniel P. Siewiorek
Carnegie Mellon University
dps@cs.cmu.edu

Asim Smailagic
Carnegie Mellon University
asim@cs.cmu.edu

Alexandre Bernardino
Instituto Superior Técnico
alex@isr.tecnico.ulisboa.pt

Sergi Bermúdez i Badia
Universidade da Madeira /
Madeira-ITI, NOVA-LINCS
sergi.bermudez@m-iti.org

ABSTRACT

Advances in artificial intelligence (AI) have made it increasingly applicable to supplement expert's decision-making in the form of a decision support system on various tasks. For instance, an AI-based system can provide therapists quantitative analysis on patient's status to improve practices of rehabilitation assessment. However, there is limited knowledge on the potential of these systems. In this paper, we present the development and evaluation of an interactive AI-based system that supports collaborative decision making with therapists for rehabilitation assessment. This system automatically identifies salient features of assessment to generate patient-specific analysis for therapists, and tunes with their feedback. In two evaluations with therapists, we found that our system supports therapists significantly higher agreement on assessment (0.71 average F1-score) than a traditional system without analysis (0.66 average F1-score, $p < 0.05$). After tuning with therapist's feedback, our system significantly improves its performance from 0.8377 to 0.9116 average F1-scores ($p < 0.01$). This work discusses the potential of a human-AI collaborative system to support more accurate decision making while learning from each other's strengths.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**; • **Applied computing** → **Health care information systems**; • **Computing methodologies** → *Artificial intelligence*; *Machine learning*.

KEYWORDS

Human-AI Interaction/Collaboration; Decision Support Systems; Explainable and Interactive Machine Learning; Personalization; Stroke Rehabilitation Assessment

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445472

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445472

1 INTRODUCTION

Advanced artificial intelligence (AI) techniques have the potential to supplement and improve decision making on various high-stake contexts (e.g. health [7, 11, 12, 35], criminal justice [26]). One promising application is a technology-assisted rehabilitation system [35, 60, 62] that supports therapists informed decision making on the assessment and administration of patients with musculoskeletal and neurological diseases (e.g. stroke). In current practices, therapists typically rely on clinical tests that involve their direct, visual observation of patient's exercise motions to evaluate the status of a patient and determine interventions [23, 45, 53]. As this process is time-consuming, therapists infrequently perform this assessment due to their limited availability [39], Thus, therapists have a lack of quantitative data on patient's performance and progress to make informed decision [6, 24]. With the goal of improving therapist's practices in rehabilitation, researchers have explored the feasibility of decision support systems for rehabilitation, which automatically monitor and assess patient's exercise motions using sensors and machine learning algorithms to generate quantitative analysis [62].

Even if prior work demonstrates the feasibility of rehabilitation monitoring systems in a laboratory setting [34], the adoption of these systems in practice still remains a challenge due to a lack of user-centered designs [7, 12, 27, 64] and the opaqueness of machine learning algorithms [10, 11, 27, 61]. These systems typically utilize labeled sensor data and a machine learning algorithm to automatically learn a function for monitoring and assessment on patient's exercises [34, 36, 62]. However, even if a complex algorithm is applied, it is challenging to derive a system that can perfectly replicate the assessment of a therapist due to diverse physical characteristics of patients. When systems with complex algorithms make incorrect predictions on assessment and do not provide any explanations on its prediction to support therapist's decision making, these black-box systems can exacerbate therapist's user experience, and be abandoned in practice [10, 11, 27, 29].

In this paper, we focus on studying how a domain expert, therapist and an interactive AI-based system can collaborate with each other on stroke rehabilitation assessment. Specifically, we develop and evaluate an interactive approach (Figure 1) that integrates a machine learning model with a rule-based model from therapists for

collaborative decision making. When a new patient performs an exercise with the patient's unaffected and affected sides, this approach first automatically selects salient kinematic features of assessment (e.g. joint angle, the trajectory of wrist to the target position, etc.) to predict the quality of motion and generate patient-specific analysis on a visualization interface (Figure 2). This patient-specific analysis includes the predicted quality of motion on three performance components (i.e. *'Range of Motion'*, *'Smoothness'*, and *'Compensation'*) and the comparison between unaffected and affected sides with the most salient features. After reviewing this patient-specific analysis, a therapist can understand the capability of a system and provide feedback (e.g. feature relevance) to refine an imperfect system.

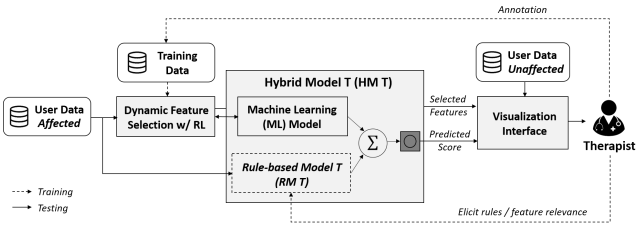


Figure 1: Flow diagram of an interactive approach for human and artificial intelligence (AI) collaborative decision making on rehabilitation assessment: an AI-based system automatically selects kinematic features of assessment to predict the quality of motion and generate patient-specific analysis on the visualization interface. A therapist can review this summarized, patient-specific analysis to improve understanding on patient's performance and provide feature-based feedback to tune a model for personalized assessment

For the implementation of our approach, we utilized the dataset of three rehabilitation exercises from 15 post-stroke and 11 healthy participants with the corresponding annotations by therapists. With this dataset, we applied reinforcement learning [37, 57] to identify the most salient features of assessment and learn a machine learning model to predict the quality of motion on patient's exercises using leave-one-patient-out cross-validation. In addition, we conducted a semi-structured interview with therapists to elicit their knowledge on stroke rehabilitation assessment into 15 independent *if-then* rules for the initial development of a rule-based model. We utilized a weighted average ensemble technique [4, 36] to combine machine learning and rule-based models into a hybrid model (HM) for assessment.

After implementing our approach, we conducted two user studies with therapists to investigate how a therapist and an AI-based system can work together on rehabilitation assessment. Our results show that therapists prefer the usage of our AI-based system with patient-specific analysis to that of a traditional system without any analysis. Specifically, the patient-specific analysis of our AI-based system empowers therapists to have richer understanding of patient's performance with quantitative measurements and supports them to achieve significantly higher agreement on assessment (i.e. 0.71 average F1-score) than the traditional system (i.e. 0.66 average F1-score) ($p < 0.05$). In addition, therapists can provide feature-based feedback to refine an imperfect AI-based system that improves its performance of replicating the therapist's assessment from 0.8377 to 0.9116 average F1-score on three exercises ($p < 0.01$).

The main contribution of this work is to present an interactive approach that supports collaboration between an expert and an AI-based system and evaluate how both an expert and an AI-based system can complement each other for more accurate decision making on rehabilitation assessment. In contrast to most related work of rehabilitation monitoring systems that focuses on improving the performance of monitoring and understanding an activity with a complex deep learning model [19, 34, 46, 49], our work highlights the importance of creating a human-centered, interactive approach for better deployment in practice. In addition, our work advances knowledge on the effect and feasibility of a human AI collaborative system for clinical decision making.

2 RELATED WORK

2.1 Towards Human-AI Collaboration

As the performance of AI systems has rapidly improved to match or exceed that of human experts [55], people have considered substituting human decision making with predictions of these systems in a variety of applications. However, deploying fully autonomous AI systems remains disruptive, dangerous, and unethical in high-stake contexts [50]. Instead, the need of involving a human to interact with AI systems has received increasing attention [1, 11, 12, 37, 59]. Amershi et al. [2] propose design guidelines for human-AI interaction through a user study with 49 design practitioners. Shneiderman [52] presents ethical principles and practical steps for human-centered AI systems. Building upon these guidelines, we focus on exploring human-AI collaboration in the high-stakes context of clinical decision making.

2.2 Human-Centered Clinical Decision Support Systems

Clinical decision support systems [44] have been considered as promising ways that can provide medical practitioners computational information on the status of a patient to improve their decision making on various disciplines (e.g. cancer diagnosis [11, 44], detection of diabetic retinopathy [7], or assessment of rehabilitation therapy [35, 60, 62]). However, even if such systems have the potential to improve the quality and efficiency of health care [44], the adoption of these systems in practice remains a challenge due to the lack of user-centered designs [3, 14, 27, 42] and the opaqueness of machine learning algorithms [10, 11, 27, 61].

For better deployment, recent research efforts in clinical decision support systems have demonstrated the value of involving the end-user in the process of design and evaluation. Yang et al. 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 [64]. Cai et al. interviewed pathologists about their desires about an AI assistant for prostate cancer diagnosis, and discussed that one major need is to make an AI assistant transparent by informing its overall capability and limitation on a task [12]. Lee

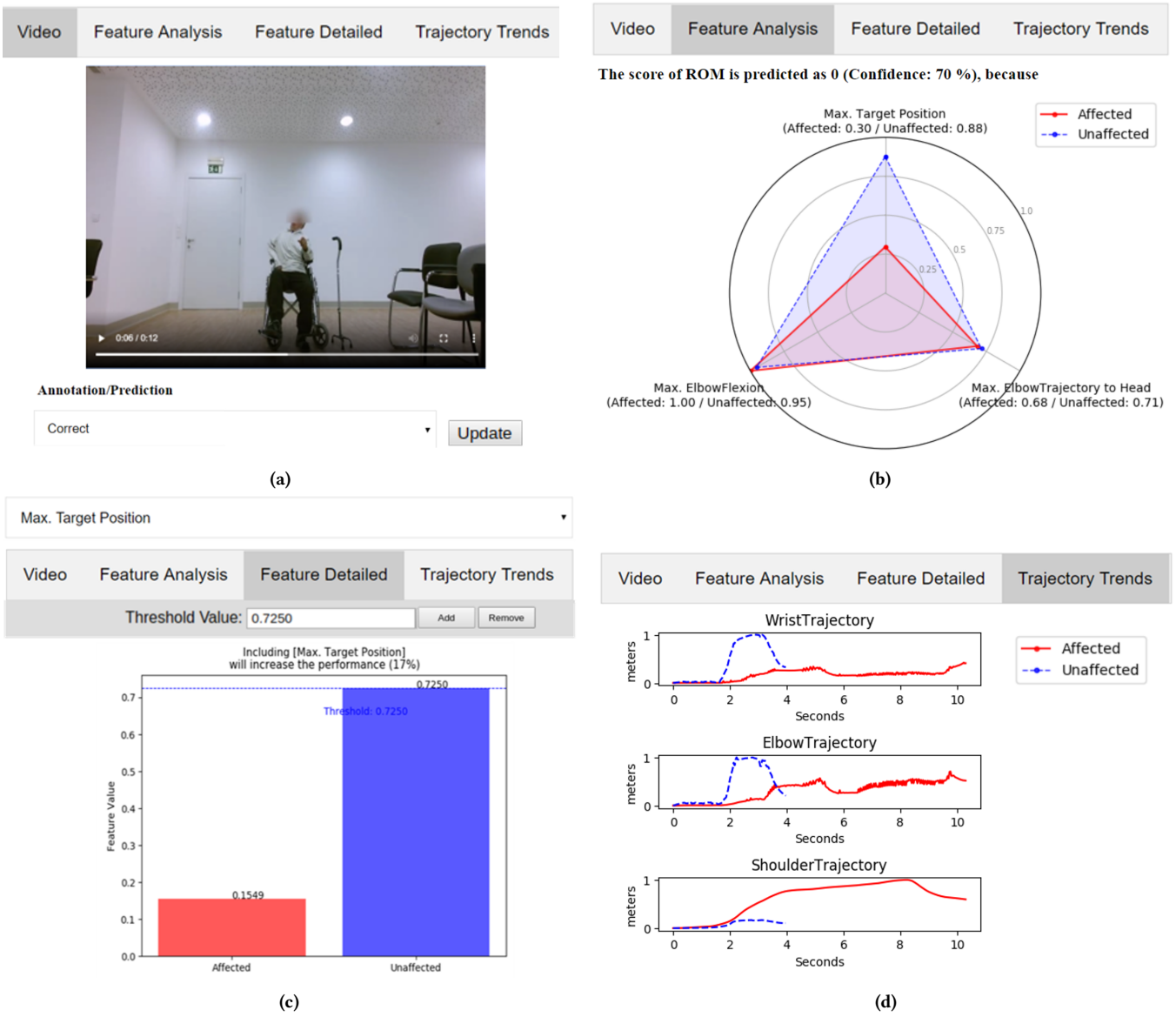


Figure 2: The visualization interface of the proposed system that presents (a) the video of patient's exercise motions and the predicted quality of motion with (b) overall feature analysis with three most important features, (c) detailed feature values with a specified threshold value of feature for assessment, and (d) trajectory trends between unaffected and affected side.

et al. [35] 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. In addition, Cai et al. demonstrated that interactive techniques can improve diagnostic utility and user trust in the content-based image retrieval systems with machine learning algorithms [11]. Beede et al. showed that several socio-environmental factors can affect the performance of a machine learning model and the experiences of practitioners and patients through an observational study in clinics [7]. These studies provide better understanding of clinicians' needs and several socio-environmental factors to deploy clinical decision support systems.

In addition, Tschandl et al. [56] presents the feasibility of AI-based supports to improve diagnostic accuracy of skin cancer recognition. Lee et al. [35] demonstrates the value of a human-centered decision support system to reduce therapists' effort on assessment and improve their agreement on assessment. However, the machine learning (ML) models of [35, 56] are fixed and provide clinicians only a passive interaction of reviewing recommendations from a system. There is still limited knowledge on human-AI collaborations, where both AI-based systems and clinicians collaborate

and complement each other on a task [12, 37]. In this work, we investigate human-AI collaboration on the clinical decision making on stroke rehabilitation assessment, and analyze the effect of an AI-based system on therapist's rehabilitation assessment and the effect of therapist's feedback on an AI-based system.

2.3 Practices and Technological-Supports for Rehabilitation Assessment

Rehabilitation assessment is a critical process for therapists to design an adequate intervention for patients with musculoskeletal and neurological diseases (e.g. stroke). Therapists typically utilize clinical tests that require direct, visual observation of patient's exercises [23, 53]. However, as these tests are time-consuming and therapists have limited availability, therapists infrequently perform assessment [39]. Thus, they have a lack of quantitative data on patient's performance and encounter a challenge of making informed decisions on patient's rehabilitation [6, 24].

To address this challenge, researchers explored the feasibility of technology-assisted systems for rehabilitation [35, 62]. These systems aim to automatically monitor and analyze motions of a patient to provide therapists quantitative insights on patient's conditions. One approach of developing these systems is to elicit a set of monitoring rules with the involvement of therapists [36]. For instance, Lee et al. compared the positions of wrist and spine joints to monitor the completion of an upper-limb exercise [36]. This rule-based approach can be easily modularized and recombined to develop a customized monitoring model. However, therapists might not be able to articulate their complex and abstract decision-making process into a set of rules. In addition, it is time-consuming to manually review abundant sensor measurements and determine which measurement could be utilized to monitor an individual status. Alternatively, a machine learning algorithm can be utilized to automatically learn a model with labeled sensor data to assess the quality of motion [34, 36, 46]. However, as the healthcare domain often involves a small dataset, it is difficult to completely replicate the therapist's assessment given patients with various conditions. In addition, when a system with a complex algorithm cannot explain its prediction to support therapist's decision making [40], therapists can lose trust on a black-box system and abandon its usage in practice [10, 27, 29, 61].

Mansoor et al. [3] discusses the necessity of more investigation to address challenges of clinical acceptance on patient monitoring systems. However, substantial prior work on technology-assisted systems for rehabilitation focuses on demonstrating the feasibility of collecting objective kinematic variables to quantify the performance of rehabilitation exercises [43, 62] and improving the performance of a machine learning model to assess the quality of motion with complex algorithms [34, 46]. Instead, our work examines the feasibility and benefits of human-AI collaborative decision making that has been noted to deserve more attention [12, 35].

2.4 Explainability & Interactive Machine Learning for Human-AI Collaboration

As an AI-based system cannot be perfect [34, 46, 55], it is critical to make the AI-based system explainable and interactive for human-AI collaborative decision making in the high-stake context. Explainability [8, 15, 16, 40] and interactive machine learning [1, 11, 32, 37] have been actively explored by researchers to create a better machine learning model with improved transparency and user acceptance. Prior work describes the value of presenting relevant information of a task and acquiring inputs of a user (e.g. constraints of a model [25], weights of features[33], or feature relevance [32, 37]) to refine the classification of a model.

This work aims to increase the interpretability of a model with feature selection [8, 28]. Specifically, this work identifies salient features for the assessment using reinforcement learning [37, 57] to predict the quality of motion and generate patient-specific analysis to summarize patient's exercise performance to a therapist. In addition, this work presents an interactive approach that integrates a machine learning model with an interpretable rule-based model to support an active engagement of therapists and make use of their knowledge [36]. After reviewing patient-specific analysis, a therapist can iteratively provide feedback on feature relevance that can be realized into a rule-based model to tune a model for personalized assessment [37]. This work contributes to increase the knowledge on how domain experts, therapists and an interactive artificial intelligence (AI) based system can augment each other for improved, collaborative decision making on rehabilitation assessment.

3 DESIGNS OF THE STUDY FOR STROKE REHABILITATION ASSESSMENT

For a test domain, this paper focuses on stroke, which is the second leading cause of death and the third most common contributor to disability [17]. After having iterative discussions with three therapists with $\mu = 5.33$, $\sigma = 2.05$ years of experience in stroke rehabilitation (TPs with check marks in the *'Specification'* column of Table 1), we specified the designs of our study on stroke rehabilitation assessment.

Table 1: Profiles of Participated Therapists for Specification, Annotation, Rule Elicitation (ElicitRule), and Feature Elicitation (ElicitFeat)

ID	Studies					# of Years in Stroke Rehab
	Specification	Annotation	ElicitRule	Evaluation	ElicitFeat	
TP1	✓	✓	✓			6
TP2	✓	✓	✓		✓	4
TP3	✓			✓		9
TP4				✓	✓	4
TP5				✓	✓	1
TP6				✓	✓	6
TP7				✓	✓	5
TP8				✓		21
TP9				✓		11

3.1 Task-Oriented Upper Limb Exercises

This study utilizes three upper-limb stroke rehabilitation exercises recommended by therapists due to their correspondence with major motion patterns [35]: elbow flexion for Exercise 1, shoulder flexion for Exercise 2, elbow extension for Exercise 3. For Exercise 1, a participant has to raise the participant's wrist to the mouth as if drinking water. For Exercise 2, a participant has to pretend to touch

a light switch on the wall. For Exercise 3, a participant has to extend the participant's elbow in the seated position to practice the usage of a cane.

3.2 Performance Components

For rehabilitation assessment, this study utilizes three performance components that are used commonly on stroke rehabilitation assessment tools (i.e. Fugl Meyer Assessment [53]) and prior works [35]. The *'ROM'* refers to the amount of a joint movement while performing a task-oriented exercise. The *'Smoothness'* describes the degree of a trembling and irregular movement on joints. The *'Compensation'* indicates whether a patient leverages unnecessary joint movements to achieve a target movement. For instance, a patient might elevate the patient's shoulder to raise the affected hand. We denote the correct/normal performance component as $Y = 1$ and incorrect/abnormal performance component as $Y = 0$.

3.3 Kinematic Features

This work represents an exercise motion with sequential joint positions from a Kinect v2 sensor (Microsoft, Redmond, USA) and extracts various kinematic features to describe performance components [34]. The features of *'ROM'* include joint angles (e.g. elbow flexion, shoulder flexion, elbow extension) and normalized relative trajectory (i.e. Euclidean distance between two joints - head and wrist, head and elbow). The *'Smoothness'* is represented by various speed-related features: the speed, acceleration, jerk, zero crossing ratio of acceleration and jerk, and Mean Arrest Period Ratio (the portion of the frames when speed exceeds 10% of the maximum speed) [51]. As our study focuses on upper-limb exercises, these speed-related features are computed on wrist and elbow joints. For the *'Compensation'*, we compute joint angles (i.e. the elevated angle of a shoulder, the tilted angle of spine, and shoulder abduction) and normalized trajectories (the distance between joint positions of head, spine, shoulder joints in x, y, z axis from the initial to the current frames).

To reduce noise of acquired joint positions from a Kinect sensor, a moving average filter with the window size of five frames is applied similar to [34]. For each exercise motion, we compute a feature matrix ($\mathbf{F} \in R^{t \times d}$) with t frame and d features of each performance component and compute statistics (i.e. max, min, range, average, and standard deviation) over all frames of the exercise to summarize a motion into a feature vector ($X \in R^{5d}$).

4 INTERACTIVE APPROACH FOR HUMAN-AI COLLABORATIVE DECISION MAKING

This work presents an interactive approach (Figure 1) that combines a machine learning (ML) model with a rule-based (RB) model from the therapist's knowledge for collaborative decision making on rehabilitation assessment. This approach first automatically identifies salient features to predict the quality of motion and generate patient-specific analysis, which supports experts to gain new insights on a decision making task [35]. After reviewing patient-specific analysis from the visualization interface (Figure 2), a therapist can provide feature-based feedback to interactively tune an imperfect model [37]. In the following subsections, we describe the components of our approach: dynamic feature selection using reinforcement learning, a machine learning (ML) model, a rule-based (RB) model, a hybrid model, and a visualization interface.

4.1 Dynamic Feature Selection using Reinforcement Learning

Reviewing kinematic features is an important way for therapists to quantitatively and objectively understand the patient's exercise performance [63]. However, simply presenting all features can overwhelm therapists and limit their ability to gain insights on the exercise performance of a patient. Given the limited availability to support multiple patients, therapists desire to minimize the amount of time on analyzing kinematic features while accurately diagnosing a patient's status. Thus, this work aims to automatically identify salient features of assessment for patient's exercise motions to generate an interpretable and succinct patient-specific analysis and improve the therapist's understanding of patient's exercise performance.

The long-established approaches of feature selection (e.g. filter, wrapper, embedded methods) [54] find a fixed feature set to the entire training dataset, which applies uniformly for all patients. In contrast, this work applies a Markov Decision Process (MDP) [37] to dynamically find the optimal feature set for each patient's motions. As each patient has a different physical and functional status [35], we hypothesize that feature selection with MDP can perform better than classical feature selection approaches for personalized rehabilitation assessment.

4.1.1 *Problem Definition.*

We formulate Markov Decision Process (MDP) for feature selection as follows:

Let $(X, Y) \in \mathcal{D}$ be a sample from a dataset, where X is a feature vector, where x_i is the value of a feature $f_i \in \mathcal{F} = \{f_1, ..., f_n\}$, n is the number of features, and Y is the class label. Let $\bar{\mathcal{F}}$ be the set of recruited features and the function $c : \mathcal{F} \rightarrow \mathbb{R}^{\leq 0}$ be the cost of adding a feature in $\mathcal{F}$.

- **State Space** ($\mathcal{S}$): Let state be $s = (X, Y, \bar{\mathcal{F}}) \in \mathcal{S}$, and the observed state of the agent be $s' = \{(x_i, f_i) \mid \forall \text{ features } f_i \in \mathcal{F}\}$, the recruited feature without the label.
- **Action Space**: Let $\mathcal{A} = \mathcal{A}_f \cup \mathcal{A}_c$ denote the action set. The agent takes either the action of selecting a feature, $\mathcal{A}_f = \mathcal{F}$, which is limited to features that are not selected or the action of classifying an instance, $\mathcal{A}_c = \mathcal{Y}$ to terminate an episode.
- **Reward**: Let the reward function be defined as
$$r(s,a) = r((X, Y, \bar{\mathcal{F}}), a) = \begin{cases} c(f_i) & \text{if } a \in \mathcal{A}_f \text{ and } a = f_i \\ -1 & \text{if } a \in \mathcal{A}_c \text{ and } a \neq Y \\ 0 & \text{if } a \in \mathcal{A}_c \text{ and } a = Y \end{cases}$$
We apply a uniform cost of selecting features: $\forall f_i,\ c(f_i) = -\lambda$, where $\lambda = 0.01$. The agent receives a reward of -1 for incorrect classification and a reward of 0 for correct classification.
- **Transition**: Let the transition function be
$$p(s,a) = p((X, Y, \bar{\mathcal{F}}), a) = \begin{cases} (X, Y, \bar{\mathcal{F}} \cup f_i) & \text{if } a \in \mathcal{A}_f \text{ and } a = f_i \\ TS & \text{if } a \in \mathcal{A}_c \end{cases},$$
where TS is the terminal state after outputting the classification and revealing the true label.

For MDP, each episode is to classify an instance from data and the environment is the power set of the feature space. An agent sequentially determines whether to query an additional feature or classify a sample while receiving a negative reward for recruiting a feature or mis-classification. We utilized the Q-network with Double Q-learning [41, 57] to solve this problem.

4.1.2 Implementation Details.
We implemented a neural network with the parameters θ (Q_θ) for Q-learning using the *'PyTorch'* library [47]. The input layer of the network consists of feature and binary mask vectors [37]. This masking input vector is to indicate whether a feature is recruited or not. Specifically, let $m \in \{0,1\}^n$ be an n-dimensional vector for an environment of n features, where $m_i = 1$ if the agent has queried feature i thus far in the episode and 0 otherwise. This target network is also used for a machine learning (ML) model to predict the quality of motion. The architecture and parameters of a neural network are described in Table 2.

For training a model, we utilize a batch of transitions that are empirically experienced by the agent with a greedy policy $\pi_\theta(s) = \text{argmax}_a\ Q_\theta(s,a)$, and apply *RMSProp* optimizer to minimize the following loss function:

$$l(\theta) = \mathbb{E}_{s,a}[(r(s,a,s') + \gamma \max_{a' \in \mathcal{A}} Q_\theta(s',a') - Q_\theta(s,a))^2] \quad (1)$$

where $r(s,a,s')$ indicates the expected immediate reward on transition from s to s' under action a and γ indicates the discounted factor. We clip a gradient if a gradient norm exceeds 1.0 and update the target network after each step. Instead of directly updating the weight of the target network, we apply soft target updates [38]: $\theta' \leftarrow \rho\theta + (1-\rho)\theta'$, where $\theta \leq 1$. ρ denotes this soft target update factor and is specified as 0.1. These soft target updates can improve the stability of learning parameters of a target network. As the application of soft target updates may lead to slow learning, we apply an experience replay [41] for sampling efficiency. Specifically, the environment with randomly drawn samples is simulated and the transition data is recorded to the experience replay buffer. As the environment is episodic with a short length, we choose a value 1.0 for the discount factor γ. In addition, we apply the ϵ-greedy policy to control the exploration. Specifically, we linearly decrease the ϵ value from the $\epsilon_{start}(0.5)$ to the $\epsilon_{end}(0.05)$ with a step value, $\epsilon_{step}(0.02)$.

4.2 Machine Learning (ML) Model

A machine learning (ML) model applies a supervised learning algorithm to predict the quality of motion on each performance component and compute the score of being correct on a performance component, $P_{ML} = P(Y = 1|X)$. This paper explores various traditional algorithms: a Decision Tree (DT), Linear Regression (LR), Support Vector Machine (SVM), a Neural Network (NN) using the *'Scikit-learn'* [48] and the *'PyTorch'* libraries [47]. For DT models, Classification and Regression Trees (CART) is applied to build prune trees while grid-searching different the maximum depth size of a tree (i.e. 3 - 5). For LR models, $L1, L2$ regularization or linear combination of $L1$ and $L2$ (ElasticNet with 0.5 ratio) are applied to avoid overfitting. For SVM models, we apply either linear, polynomial, or Radial Basis Function (RBF) kernels with penalty parameters, $C = 1.0$. For NN models, we grid-search various architectures (i.e. one to three layers with 32, 64, 128, 256, 512 hidden units) and different learning rates (i.e. 0.0001, 0.005, 0.001, 0.01, 0.1) and apply the *'ReLu'* activation functions and *'AdamOptimizer'*. NN models are trained until the tolerance of optimization is 0.0001 or the maximum 200 iterations.

4.3 Rule-based (RB) Model

A rule-based (RB) model leverages a set of feature-based rules from therapists to predict the quality of motion. For the initial implementation, we conducted a semi-structured interview with two therapists to elicit their knowledge of assessing stroke rehabilitation exercises. This knowledge is represented as 15 independent *if-then* rules. For example, assessing ROM for Exercise 1 is defined as follows [36]:

$$\hat{Y} = \begin{cases} 1 & \text{if } p^{max}(wr, c_y) >= p^{max}(spsh, c_y) \\ 0 & \text{else} \end{cases}$$

where $p(j,c)$ indicates a joint position with a joint j (e.g. the wrist (wr) and the spine shoulder, the top of spine, ($spsh$)) and the coordinate of a joint, c in the set $C \in \{c_x, c_y, c_z\}$. $\hat{Y}$ denotes the predicted label on a performance component. This rule compares the maximum position of the wrist joint, $p^{max}(wr, c_y)$ with that of the spine shoulder joint $p^{max}(sh, c_y)$ in the y-coordinate to roughly estimate whether a patient achieves the target position of the Exercise 1, *'Bring Cup To the Mouth'*.

A rule-based (RB) model computes the score of being correct on the performance component as follows:

$$P_{RB^T} = \frac{1}{|\mathbb{R}^T|} \sum_{r \in \mathbb{R}^T} \min(\frac{f_r}{\tau_r}, 1) \quad (2)$$

where f_r describes the feature value of a rule r from an exercise motion (e.g. $p^{max}(wr, c_y)$ for the example above) and τ_r describes the threshold value of a rule r (e.g. $p^{max}(spsh, c_y)$ for the example above). $\mathbb{R}^T$ indicates the set of rules from the therapists with T-th iteration. min function is applied so that this equation assigns a value of 1 if the feature value of a rule exceeds the threshold of that rule. Otherwise, the equation normalizes the feature value of a rule with the threshold of a rule to compute the likelihood of being correct.

A rule-based (RB) model can be iteratively updated with the therapist's feature-based feedback for personalized assessment with patient-specific rules. Our approach can identify salient features of assessment for patient's motions and generate predicted assessment with a patient-specific feature analysis between patient's unaffected and affected motions (Figure 2b): *'Incorrect'* ROM is predicted due to smaller maximum target position (affected: 0.30, unaffected: 0.88), maximum elbow flexion (affected: 1.00 / unaffected: 0.95), and maximum elbow trajectory to head (affected: 0.68 / unaffected: 0.71). A therapist can review this patient-specific analysis and provide feedback whether identified, presented features should be included or excluded [32] to predict assessment (Figure 2c). When including a feature into a rule-based model, a therapist can either specify a threshold value or utilize the feature value of the unaffected side for the threshold value of a feature-based rule (τ_r). Feedback of adding

or removing a feature indicates that the corresponding feature-based rule will be included or removed from the current set of rules ($\mathbb{R}^T$). Similarly, feedback of updating a threshold value indicates to replace the threshold value (τ_r) of the corresponding rule in the current set of rules ($\mathbb{R}^T$).

4.4 Hybrid Model

A hybrid model (HM) utilizes a weighted average, ensemble technique [4, 36] to combine two perspectives on assessment: a data-driven, machine learning (ML) model and a rule-based (RB) model from therapists. For the classification of the quality of motion, the HM computes a weighted average of prediction scores from two models, in which the contribution of each model is weighted by the performance of a model (i.e. the F1-score of each model in the range of [0, 1]). The prediction score of the HM, P_{HM} is computed as follows:

$$P_{HM^T} = \frac{\rho_{ml}}{\rho_{ml} + \rho_{rb^T}} P_{ML} + \frac{\rho_{rb^T}}{\rho_{ml} + \rho_{rb^T}} P_{RB^T} \quad (3)$$

where P_{ML} and P_{RB^T} indicate the predicted scores of a machine learning (ML) model and a rule-based (RB) model with T-th iteration from therapists and ρ_{ml} and ρ_{rb^T} describe the performance, F1-scores of a ML model and a RB model with T-th iteration respectively.

4.5 Visualization Interface

Based on a few guidelines of Human Artificial Intelligence (AI) interaction [2, 32], we designed and implemented the web-based visualization interface. This interface shows a video of recorded patient's exercise motions (Figure 2a). As therapists desire quantitative feature analysis instead of repetitively watching a video patient's exercise motions [35], this interface also presents a patient-specific analysis that is considered *"contextually relevant information"* [2] for the assessment. This patient-specific analysis includes the predicted quality of each performance component (e.g. *'Range of Motion', 'Smoothness', 'Compensation'*) and supplementary information on the prediction of a model, which contains feature analysis (Figure 2b and 2c) and trajectory trends (Figure 2d) [35]. When presenting the predicted quality of motion on performance components, the performance of a system is also included to *"make clear how well the system can do"* [2].

In practice, therapists utilize patient's unaffected motions as normality for assessment [45]. To follow this practice, *"social norms"* [2], the interface includes comparison between the affected and unaffected side of a patient to present salient features and trajectory trends of three major joints (e.g. shoulder, elbow, and wrist) for upper-limb exercises [35]. The interface *"avoids overwhelming"* [32] therapists by including only three salient features for each performance component with highest information gain. A radar chart [35] is utilized to present multivariate, kinematic features effectively.

In addition, the interface supports to *"honor user feedback"* [32]: feature-based feedback from a therapist. After reviewing patient-specific feature analysis, a therapist can provide feature-based feedback on whether an identified feature on analysis should be included or removed for assessment (Figure 2c). For including a feature, a therapist can specify a patient-specific threshold value to generate a feature-based rule for personalized rehabilitation assessment. This interface presents a change in the performance of a model to support the elicitation on the therapist's feature-based feedback (e.g. *"Including Max. Target Position will increase the performance (17%)"* in Figure 2c).

5 SYSTEM IMPLEMENTATION STUDY

5.1 Dataset of Stroke Rehabilitation Exercises

We utilized the dataset of three upper-limb stroke rehabilitation exercises from 15 post-stroke (13 males and 2 females) and 11 healthy (10 males and 1 female) participants using a Kinect v2 sensor (Microsoft, Redmond, USA). During data collection, a sensor was located at a height of 0.72m above the floor and 2.5m away from a participant and recorded the 3D trajectory of joints and video frames at 30 Hz. The starting and ending frames of exercise movements were manually annotated.

A post-stroke patient participated in two data collection sessions. In the first session, a therapist evaluated post-stroke patient's functional ability using the a clinically validated tool, Fugl Meyer Assessment (FMA) [53]. Post-stroke patients had diverse functional abilities from mild to severe impairment (37 ± 21 Fugl Meyer Scores out of 66 points). In the second session, post-stroke patients performed 10 repetitions of each exercise with both their affected and unaffected sides. Eleven healthy participants engaged in a single session, which each participant performed 15 repetitions with participant's dominant arms for each exercise.

Two therapists (TP 1 and 2 with check marks in the *'Annotation'* column of Table 1) individually annotated the dataset to implement our approach and compute the agreement level of therapists. They separately watched the recorded videos of the patient's exercises (Figure 2a) and annotated the performance components of the patient's exercises without reviewing the analysis of our system (Figure 2b, 2c, 2d).

5.2 Evaluation Methods

For implementation, the collected data is divided into *'Training'* and *'User'* data:

- ***'Training Data'*** (Figure 1) is composed of 165 unaffected motions from 11 healthy participants and 140 affected motions from 14 stroke patients to train a feature selection model and a machine learning (ML) model.
- ***'User Data'*** (Figure 1) includes held-out testing post-stroke patient's unaffected and affected motions. Given testing the patient's affected motions, our approach dynamically selects salient features of assessment and predicts the quality of motion on performance components. For feature analysis of the visualization interface (Figure 2b and 2c), both unaffected and affected motions of a testing patient are utilized.

To train a machine learning (ML) model, we utilized the annotation of therapist 1 (TP 1), who had more interactions with recruited post-stroke patients by evaluating their functional ability with the Fugl Meyer Assessment [53]. Given this ground truth annotation, we applied leave-one-patient-out (LOPO) cross validation on post-stroke patients to implement and evaluate our feature selection and

machine learning models. During LOPO cross validation, a model was trained with data from all participants except one post-stroke patient and tested with affected motions of the left-out post-stroke patient. This process was repeated to evaluate affected motions of all post-stroke patients. For the performance metric, we utilized a F1-score that seeks to balance between precision (i.e. how many instances a model can classify correctly) and recall (i.e. how robust a model is). The F1-score can provide a more realistic measure of a model and be beneficial when there is an uneven class distribution.

To analyze performance of machine learning models, we compared the annotations of therapist 1 and 2 to compute the agreement level of therapists with F1-scores (Figure 3), In addition, we conducted paired t-tests over three performance components of three exercises to compare performance of our approach with various machine learning (ML) models.

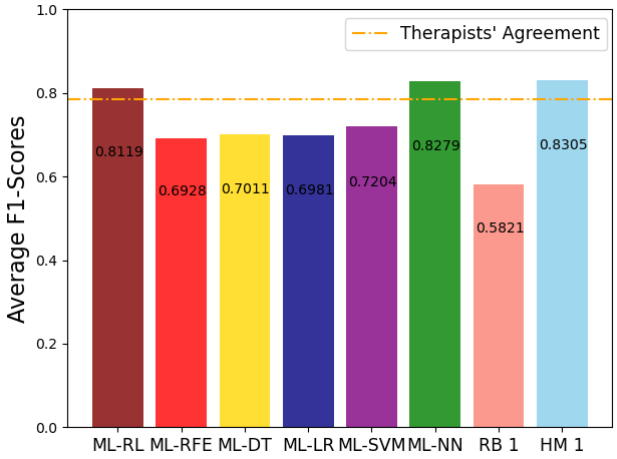


Figure 3: Performance (average F1-scores) of various machine learning (ML) models, non-interactive, initial rule-based and hybrid models (RB 1 and HM 1) and therapists' agreement.

5.3 System Implementation Results

Figure 3 summarizes the performance of various models, which measures the agreement with the therapist's annotation using an average F1-score over performance components of three exercises. The parameters of machine learning models that achieve the best F1-score during Leave-One-Patient-Out (LOPO) cross-validation are summarized in the Table 2.

For machine learning (ML) models, we present the performance of neural network trained for feature selection using reinforcement learning (ML-RL), feature selection using Recursive Feature Elimination (ML - RFE), one of classical feature selection methods [20], a decision tree (ML-DT), linear regression (ML-LR), a support vector machine (ML-SVM), and a neural network trained with the full set of features (ML-NN). Our approach, ML-RL achieves decent agreement with the therapist's annotation: 0.8119 average F1-score over three exercises. In terms of feature selection approaches, ML-RL has 0.11 higher average F1-score than ML-RFE ($p < 0.01$) and is expected to perform better to generate a patient-specific analysis for therapists. Compared to ML-NN with the full set of features, ML-RL has 0.016 lower F1-score. However, ML-RL still outperforms machine learning models with other algorithms: decision tree (ML-DT with 0.7011 average F1-score), linear regression (ML-LR with 0.6981 average F1-score), and support vector machine (ML-SVM with 0.7204 average F1-score) ($p < 0.01$). The performance of ML-RL is equally good with that of the therapists' agreement.

The initial, non-interactive rule-based model (RB 1) achieves the lowest agreement level with the therapist's annotation: 0.5821 average F1-score over three exercises. For the initial, non-interactive hybrid model (HM 1), we combine the machine learning model with feature selection using reinforcement learning (ML-RL) with the initial rule-based model (RB 1). The HM 1 achieves 0.8305 average F1-score over three exercises, which is significantly better performance than machine learning models with decision trees (ML-DT), linear regression (ML-LR), and support vector machine (ML-SVM) ($p < 0.01$). However, the performance of the HM 1 is not significantly different, equally good with that of ML-NN and the therapists' agreement (TPA). Integrating two models of assessment does not significantly improve the performance of a model.

6 USER STUDY ON HUMAN-AI COLLABORATIVE DECISION MAKING

After the system implementation study, we conducted user studies with therapists to explore how a therapist and an interactive artificial intelligence (AI) based system can collaborate on rehabilitation assessment. Specifically, we investigate 1) the effect of patient-specific analysis from an AI-based system (i.e. predicted quality of motion on performance components, feature analysis, and trajectory trends) on a therapist's decision making process during rehabilitation assessment and 2) the effect of accommodating therapist's feedback on an AI-based system. The questions our studies sought to answer were:

- RQ 1: How does patient-specific analysis from an AI-based system affect the therapist's experience during decision making on patient's rehabilitation assessment?
- RQ 2: Can a therapist provide feedback on an AI-based system to improve its performance?

6.1 Study to Evaluate the Utility of an AI-based System

Although prior work demonstrates the feasibility of developing an AI-based system for rehabilitation monitoring and assessment [34, 62], limited work explores how such a system is actually used in practice and affects user's experiences on decision making [9]. To evaluate the effect of an AI-based system, we compared the therapist's experience on assessing post-stroke patient's exercises using our proposed system (Figure 2) to two baseline systems: the *'Traditional'* system that presents only videos for assessment and the *'Predicted Scores (PredScore)'* system that presents videos with predicted scores without any patient-specific analysis. In the study, we referred to systems as "Condition 1", "Condition 2", and "Condition 3" (counterbalanced) to avoid biasing participants. For clarity,

we refer to them as the "proposed", "traditional", and "predicted score (predscore)" systems in this paper.

6.1.1 Metrics.
We evaluated the systems with respect to 1) responses on questionnaires from therapists and 2) their agreement level on assessment. Questionnaires are utilized [11] to collect therapist's opinions on various aspects of a system: usefulness and richness to support decision making on assessment and attitudes toward an system (e.g. trust, workload, usage intention). All questionnaires were rated on a 7-point scale.

- Usefulness: *"[System - Condition X] is useful to understand and assess patient's performance"*
- Richness: *"[System - Condition X] generates new insights on patient's performance"*
- Trust: *"I can trust information from [System - Condition X]"*
- Workload: participants answered the *"effort"* and *"frustration"* dimensions of the NASA-TLX [21] (e.g. *"How hard did you have to work to accomplish the evaluation task using the [System - Condition X]?"* and *"How insecure, discouraged, imitated, stressed were you while using the [System - Condition X]?"*
- Usage Intention: *"I would use [System - Condition X] to understand and assess patient's performance"*

In most medical disciplines, experts can be biased in their decision making based on their own experiences and expert disagreement is prevalent [5, 30, 31], even if they rely on standardized guidelines [18, 23, 53]. Thus, this study also utilizes the agreement level of therapists' assessment to analyze the effect of a decision support system. During the study, we collected the therapist's assessment on the patient's rehabilitation exercises while using each system. With this assessment from therapists, we computed their agreement level on assessment to analyze whether our proposed system with patient-specific analysis supports them more consistent assessment than two baseline systems.

6.1.2 Procedures.
Seven therapists with $\mu = 8.14$, $\sigma = 6.05$ years of experience in stroke rehabilitation (TPs with check marks in the *'Evaluation'* column of Table 1) from four rehabilitation centers participated in the user study on the evaluation. Note that we excluded two therapists (TP 1 and 2), who annotated the dataset to implement our system. After obtaining IRB approved informed consent, each therapist was instructed on the procedure of the study and systems using dummy data (30 minutes). Then, a therapist was assigned to the task of assessing videos (around one minute per video, in which a patient performs a rehabilitation exercise) using each system and followed by post-study questionnaires (1.5 hours total).

We assigned the task of assessing 15 videos (five patients performing three exercises) on each system. To counterbalance the task of assessing videos on each system, therapist 1 (TP 1), who evaluated the functional ability of patients, divided patients into three sub-groups, in which patients of each subgroup have similar functional abilities. The order of presenting the conditions/systems and the assignment of a subgroup on a system are randomized. After finishing a task using each system, therapists responded to the questionnaires on the corresponding system. When therapists completed tasks with all systems, they participated in post-interviews to further describe their perspectives on the effectiveness of using our system during the assessment.

6.2 Study to Collect Therapist's Feedback on an AI-based System

Our interactive AI-based system (Figure 1) allows a therapist to play an active engagement for collaborative decision making with the system instead of providing a passive interaction (e.g. reviewing the prediction of a system). To analyze the effect of the therapist's feedback on an AI-based system, we had an additional study with five therapists with $\mu = 4.0$, $\sigma = 1.67$ years of experience in stroke rehabilitation (TPs with check marks in the *'ElicitFeat'* column of Table 1). Each therapist was instructed on the task of providing feature-based feedback with dummy data. The task of a therapist is to review patient-specific analysis with predicted assessment and salient features (Figure 2b and 2c) and provide feedback to make the predicted assessment from the system as accurate as possible during a 30 minutes session. For feature-based feedback, a therapist has the following three options: 1) include or 2) remove a selected feature for assessment, or 3) update the threshold value of a selected feature for assessment. We assigned non-overlapping, three patients for each therapist to generate feature-based feedback on all post-stroke patients in our dataset.

7 USER STUDY RESULTS

7.1 Effect of an AI-based System on Therapist's Assessment

For analysis of results, we first performed one-way ANOVA tests on results of metrics (i.e. responses on questionnaires from therapists and their agreement level on assessment). If the results have any statistical significance, we further performed pairwise statistical analysis on three conditions/systems using paired t-tests. The results of statistical analysis using both one-way ANOVA tests and post-hoc, paired t-tests are summarized in Table 3.

Figure 4 describes the responses on questionnaires from participated therapists and their agreement level during rehabilitation assessment. Overall, the proposed system has received better responses on all questionnaires: it presents therapists more useful and richer information, reduces their efforts and frustration on stroke rehabilitation assessment, and achieves the highest score on usage intention. However, when we analyzed the statistical significance using one-way ANOVA tests and post-hoc, paired t-tests, only **richness, effort**, and **usage intent** variables have statistical significance. Specifically, the proposed system has a **significantly higher richness** score ($\mu = 6.00$) than the others (traditional: $\mu = 5.06, p < 0.10$ and proposed: $\mu = 4.83, p < 0.10$). The scores of richness between traditional and predscore systems have no statistical difference (Table 3), this indicates the positive effect of user-specific analysis from the proposed system on the richness. Therapists experienced **significantly lower effort** on assessment ($\mu = 2.66$) with the proposed system than the others (traditional: $\mu = 3.93, p < 0.10$ and predscore: $\mu = 3.11, p = 0.27$). The effort score of the predscore system does not have a statistical difference

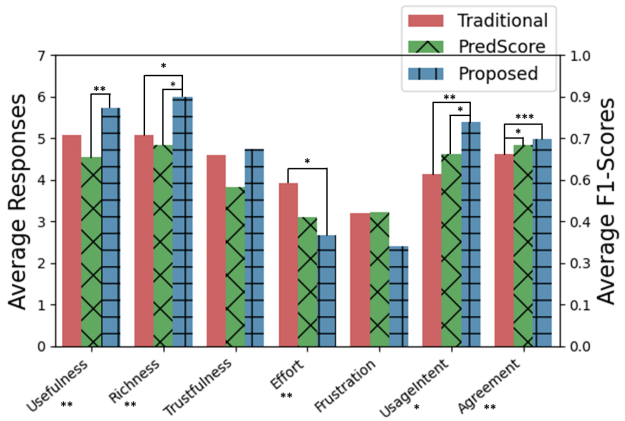


Figure 4: Results of the User Study in term of (a) Responses of Questionnaires and (b) Agreement Level of Therapists' Assessments: the proposed system is more useful, richer, more trustful while reducing effort and frustration on assessment tasks. It is more likely to be used in the clinical practices than other baseline systems (i.e. traditional and predscore), and achieves the highest agreement on therapists' assessment. We indicate statistical significance using one-way ANOVA tests at the bottom of each variable and post-hoc, paired t-tests at the top of each variable. *, **, and * indicate statistical significance using the one-way ANOVA tests or paired t-tests at 90%, 95%, and 99% significance level.**

with that of the traditional system (Table 3). Thus, user-specific analysis of the proposed system has a positive effect on lower effort. In addition, the proposed system has a **significantly higher usage intent** score ($\mu = 5.4$) than the others (traditional: $\mu = 4.13$, $p < 0.05$ and predscore: $\mu = 4.61$, $p < 0.10$). As the usage intention scores between traditional and predscore systems are not statistically different (Table 3), the user-specific analysis of the proposed system has a positive effect on the usage intent score.

When we further analyze therapists' assessment from three systems, the proposed system with predicted assessment and user-specific analysis (i.e. feature analysis, salient frames, and graphs of joint trajectories) has supported therapists to achieve **significantly higher agreement** on assessment ($\mu = 0.7138$ F1-score) than the others: the traditional system ($\mu = 0.66$ F1-score , $p < 0.01$) and the predscore system ($\mu = 0.6924$ F1-score, $p = 0.18$). Although both the predscore and proposed systems achieve higher agreement levels than the traditional system, the difference between the traditional and proposed systems ($p < 0.01$) has higher statistical significance than the difference between the traditional and predscore systems ($p < 0.10$). Thus, this indicates a more positive effect of user-specific analysis from the proposed system to improve the agreement level of therapists' assessment.

According to comments from therapists, presentation of predicted assessment and patient-specific analysis from the proposed system is considered *"useful to understand patient's condition"* and *"validate my own assessment"*. Even though some predictions from the proposed system are *"not matching and trustful"*, patient-specific analysis of the proposed system complements to *"understand why such predicted scores are generated"*. Thus, therapists can understand the competence of a system to predict assessment on patient's exercises and develop a mental model in which cases therapists can trust a system or not. In addition, therapists described that patient-specific analysis of the proposed system (feature analysis and trajectory trends) helps reduce their effort and frustration to *"find evidence for assessment in videos"*. They considered the proposed system as *"a good platform"*, which *"brings more interesting, new aspects of a patient"* in an easy and intuitive way and could facilitate *"understanding on patient's performance and communicate it with patients"*. Overall, therapists are positive to accept the usage of our proposed system in practice.

7.2 Effect of Therapist's Feedback on an AI-based System

After reviewing assigned patient-specific analysis from our system, therapists provided nine feature-based feedback on each patient to tune a system for rehabilitation assessment. Specifically, therapists provided 7.26 new features, removed 0.33 features, and updated 1.06 threshold values on average over 15 post-stroke patients.

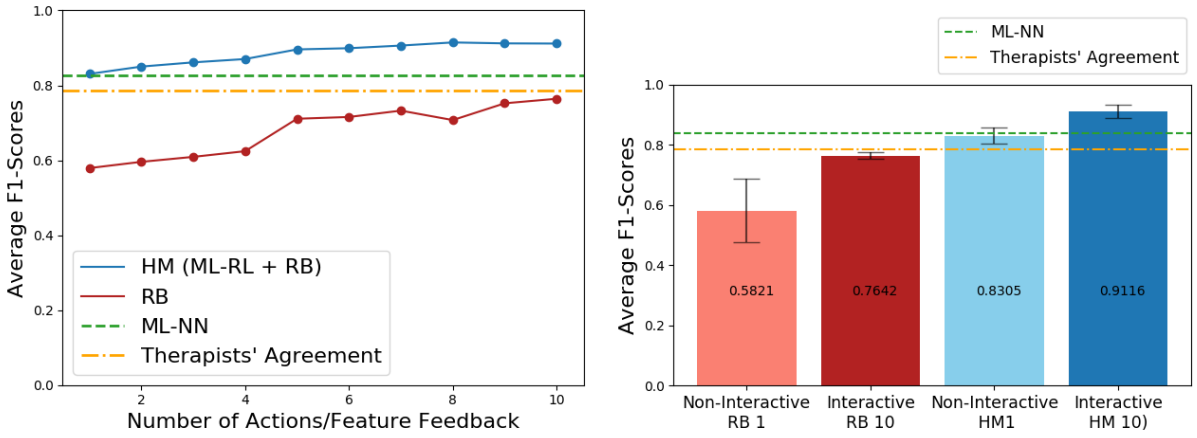
Both the rule-based (RB) model and hybrid model (HM) significantly improve their performance while accommodating the therapist's nine feature-based feedback on each patient (Figure 5). Specifically, the rule-based (RB) model significantly improves its performance to replicate the therapist's assessment, 31% from 0.5821 to 0.7642 average F1-scores over three exercises ($p < 0.01$). The tuned, interactive RB model (RB 10) has 0.0437 higher average F1-score than the machine learning model with Support Vector Machine (ML-SVM in Figure 3) and comparable performance with the therapists' agreement (i.e. 0.0212 lower average F1-score). In addition, the hybrid model (HM) also significantly improves its performance to replicate the therapist's assessment, 9.7% from 0.8305 to 0.9116 average F1-scores ($p < 0.01$). The tuned, interactive HM (HM 10) outperforms both the machine learning model with Neural Networks (ML-NN in Figure 3) and therapists' agreement (Figure 3): 0.0739 higher average F1-score than ML-NN and 0.1262 higher average F1-score than therapists' agreement ($p < 0.01$).

8 DISCUSSION

In this work, we study and discuss how a domain expert and an artificial intelligence (AI) based system can collaborate for a decision making task on stroke rehabilitation assessment.

8.1 Improving Experts' Decision Making with an AI-based System

Most medical disciplines rely on standardized guidelines to support expert's decision makings [18, 53]. However, as these guidelines are limited to high-levels, experts can still become uncertain about applying these guidelines and biased in their decision making based on their experiences [10, 13], and expert disagreement is prevalent [5, 30, 31]. Our results demonstrate that an artificial intelligence (AI) based system can provide therapists quantitative insights on the status of a patient to improve their experiences and agreement level of rehabilitation assessment. Instead of presenting abundant



(a) (b)

Figure 5: The effect of therapist's feature-based feedback: (a) the performance of models over feedback, iterations and (b) the comparison of model performance without/with feature-based feedback. Both rule-based and hybrid models significantly improve their performance with therapist's feature-based feedback ($p < 0.01$ using paired t-tests). Interactive, tuned hybrid model (HM 10) performs better than the machine learning model with neural networks (ML-NN) and therapists' agreement.

quantitative data for expert's review, our system can automatically identify salient kinematic features of decision making on rehabilitation assessment to predict assessment and generate succinct patient-specific analysis as explanations on its prediction.

Therapists considered that the patient-specific analysis of our system is useful to understand the performance of a patient quantitatively, especially when a patient performs a motion incorrectly. For instance, when a patient partially achieved the target position of an exercise, a therapist had to watch a video of the patient's exercise repeatedly and derive an imaginary threshold boundary to distinguish whether the patient's motion is closer to the half of the target motion or not. This granular assessment from watching a video becomes *"more challenging and complex when the performance of a patient is on the edge of two classes of the assessment"*. In contrast to a traditional system without analysis, our system with patient-specific analysis assisted therapists to experience lower effort to find evidence for assessment (Figure 4) and validate their observation-based assessment with objective data. Even if a system did not always provide a prediction that matched with therapist's assessment, therapists commented that *"patient-specific analysis with salient kinematic features still brought interesting aspects on the assessment (e.g. comparing difference on unaffected and affected side quantitatively and observing the trends of a joint trajectory)"*. They felt more assured on their assessment after reviewing this patient-specific analysis with objective data.

Therapists could also develop a mental model on the competency of a system [12, 35] while validating their decision with patient-specific analysis. Based on this mental model, therapists trusted and utilized a system accordingly, but also generated an idea on how an imperfect system should be improved. For instance, therapists found that the predictions on the *'ROM'* performance component matched well with their assessment than those on the *'Compensation'* performance component. They mentioned that they spent more time on analyzing patient-specific analysis of the *'Compensation'* performance component than that of the *'ROM'* performance component. After reviewing patient-specific analysis, they selectively utilized information from a system on their decision making. In addition, therapists commented on the possibility of interacting with a system to refine a system. *"I found a case when a system misclassified compensation motion that involved leaning trunk forward, but it did not present a leaning trunk forward feature in the analysis. It could be better if I can interact with a system and adjust feature selection of a system over time"*.

Overall, the patient-specific analysis of our AI-based system supported therapists to achieve more consistent assessment on patient's exercises than the traditional interface without any patient-specific analysis. Most therapists preferred the usage of our system in practice instead of repeatedly watching videos during the assessment. Our findings highlight the importance of presenting supplementary, quantitative information on the prediction of a system for more accurate decision making and the potential value of an interactive approach that allows a therapist to tune an imperfect system for better acceptance in practice.

8.2 Improving an Imperfect AI-based System with Expert's Feedback

Artificial intelligence (AI) based systems continue to improve their performance on various tasks [34, 55, 58]. However, such AI systems may not completely replicate experts' knowledge in a healthcare domain that often involves a small dataset and may not perform well on unobserved examples. Instead of relying on only machine learning algorithms, we designed and implemented an interactive approach (Figure 1) that combines a machine learning model with a rule-based model from therapists to support collaborative decision making on stroke rehabilitation assessment.

Among various non-hybrid models, machine learning models with Neural Networks (ML-NN in Figure 3) outperform other non-hybrid models and achieves comparable, equally good performance with the initial hybrid model (HM 1 in Figure 3) that integrates the machine learning model with reinforcement learning-based feature selection (ML-RL) with the initial rule-based model (RB 1). However, the ML-NN has the limitation of being a black-box model, in which therapists could only have a passive interaction (e.g. reviewing the output of a model). Therapists could lose trust on such a black-box model with a passive interaction, and abandon its usage in practice [10, 27, 29, 61]. In contrast to the ML-NN, the HM has the benefit of supporting a therapist with the interpretation and refinement of a model by analyzing and updating rules of a rule-based model.

We further explored the possibility of interacting with our system to adjust feature selection and rule-based model of a system based on therapists' comments (Section 8.1). Our results show that this interactive approach with a hybrid model provides experts an opportunity to actively engage with the outputs of a model and iteratively provide feedback on an imperfect model for personalized assessment with improved performance. After reviewing the patient-specific analysis, therapists can have a better understanding of the patient's exercise performance and the capability of a model, and provide feature-based feedback to refine a model. As initial high-level rules from therapists are not tuned for each patient, the initial rule-based model (RB 1 in Figure 3) has the lowest performance. After tuning with the feature-based feedback from therapists, the rule-based (RB) model significantly improves its performance that is comparable to the machine learning model with Support Vector Machine (ML-SVM in Figure 3). This tuned, rule-based model (RB 10) provides another perspective on assessment, which also leads to significant improvement on the performance of the hybrid model (Figure 5b). The tuned, hybrid model (HM 10) achieves significantly better performance than ML-NN and therapists' agreement (Figure 5b). Even if an AI-based system still cannot replicate perfectly the therapist's assessment, our interactive approach supports expert's active engagement for better acceptance in practice [1, 11].

8.3 Potential Clinical Impact and Limitations

During a clinical decision-making task, both an expert and an artificial intelligence (AI) based system encounter a challenge respectively. Experts need to make a decision under uncertainty based on standard guidelines and their own experiences [10, 13, 53]. An AI-based system can suffer to learn expert's decision making from a small dataset [22, 34]. Our findings show that both experts and an AI-based system can learn from each other's strength and make more accurate, collaborative decision making. An AI-based system can serve as an assistant of experts [12, 35] to provide new quantitative insights to improve the expert's decision-making process. After interacting with an AI-based system, an expert can understand the capabilities and limitations of a system [12, 35] and provide feedback on a system for improvement. We believe that general concepts of our interactive approach that present predicted expert's decision making, explain its predictions on decision making with salient features, and accommodate expert's feedback can be applicable to other disciplines for improving decision making. However, this study is limited to demonstrate on a single domain, stroke rehabilitation assessment with feature-based feedback from a few interactions with therapists and ground truth labels of a single therapist. More investigations on ground truth labels of multiple therapists, different tasks, or deployment over an extended period are needed for further generalization.

As each discipline and decision making rely on different data modalities and priorities, it might be difficult to have a unified way of interactions between an expert and an AI-based system on various decision making tasks. For instance, our work utilizes sensor measurements on the positions of body joints in contrast to the prior work on decision support systems with image data [7, 11, 12]. In case of using sensor measurements, it is important to confirm these measurements with domain experts and provide these measurements in terms of understandable and user-friendly terminologies [35]. In addition, as refinement tools on image data (e.g. emphasizing or de-emphasizing an image) [11] are not applicable for sensor measurements from body joint positions, we designed and implemented a feature-based rule as a way to support therapist's engagement with an AI-based system. Thus, we recommend a human centered design process [7, 11, 35] on each decision making task to specify the customized forms of explanations and interactions with an AI system for human-AI collaborative decision making in practice.

9 CONCLUSION

In this paper, we described the implementation and evaluation of an interactive approach for human and artificial intelligence (AI) collaborative decision making. Our results show that both a domain expert and an AI-based system can learn from one's strength over the interaction and generate collective, hybrid intelligence on a complex decision making task with improved accuracy. An AI-based system with feature selection can provide experts summarized quantitative analysis (e.g. predicted decision making and explanations on its prediction with salient features) and support them more consistent decision making. After interacting with an AI-based system, experts can understand the limitation of a system and provide feedback to improve an imperfect system. This work contributed to broaden and enrich knowledge on how a domain expert and an AI based system can collaborate with each other on a complex decision making task (e.g. stroke rehabilitation assessment).

ACKNOWLEDGMENTS

The authors thank all the participants in this study for their dedication, time, and valuable inputs. Also, we thank the anonymous reviewers for their constructive comments on the manuscript. This work is partially supported by the National Science Foundation (NSF) under grant number CNS-1518865. Additional support was provided by the IntelligentCare project (LISBOA-01-0247-FEDER-045948), co-financed by the ERDF through the LISBOA 2020 and the FCT under CMU-PT and the FCT LARSyS - Plurianual funding 2020-2023 (UIDB/50009/2020).

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 3.

[3] Mirza Mansoor Baig, Hamid GholamHosseini, Aasia A Moqeem, Farhaan Mirza, and Maria Lindén. 2017. A systematic review of wearable patient monitoring systems–current challenges and opportunities for clinical adoption. *Journal of medical systems* 41, 7 (2017), 115.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443.

[5] 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), e190096–e190096.

[6] Mark T Bayley, Amanda Hurdowar, Carol L Richards, Nicol Korner-Bitensky, Sharon Wood-Dauphinee, Janice J Eng, Marilyn McKay-Lyons, Edward Harrison, Robert Teasell, Margaret Harrison, et al. 2012. Barriers to implementation of stroke rehabilitation evidence: findings from a multi-site pilot project. *Disability and rehabilitation* 34, 19 (2012), 1633–1638.

[7] 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.

[8] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 1.

[9] Tiffani J Bright, Anthony Wong, Ravi Dhurjati, Erin Bristow, Lori Bastian, Remy R Coeytaux, Gregory Samsa, Vic Hasselblad, John W Williams, Michael D Musty, et al. 2012. Effect of clinical decision-support systems: a systematic review. *Annals of internal medicine* 157, 1 (2012), 29–43.

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

[11] 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, 4.

[12] 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.

[13] Wendy J Coster. 2008. Embracing ambiguity: Facing the challenge of measurement. *American Journal of Occupational Therapy* 62, 6 (2008), 743–752.

[14] Yaron Denekamp. 2007. Clinical decision support systems for addressing information needs of physicians. *The Israel Medical Association journal: IMAJ* 9, 11 (2007), 771–776.

[15] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[16] Mengnan Du, Ninghao Liu, and Xia Hu. 2018. Techniques for interpretable machine learning. *arXiv preprint arXiv:1808.00033* (2018).

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

[18] 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.

[19] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition Using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 11 (June 2017), 28 pages. https://doi.org/10.1145/3090076

[20] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning* 46, 1-3 (2002), 389–422.

[21] 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.

[22] Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

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

[24] Mark Jones, Karen Grimmer, Ian Edwards, Joy Higgs, and Franziska Trede. 2006. Challenges in applying best evidence to physiotherapy. *Internet Journal of Allied Health Sciences and Practice* 4, 3 (2006), 11.

[25] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1343–1352.

[26] Danielle Leah Kehl and Samuel Ari Kessler. 2017. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. (2017).

[27] 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.

[28] Been Kim, Julie A Shah, and Finale Doshi-Velez. 2015. Mind the gap: A generative approach to interpretable feature selection and extraction. In *Advances in Neural Information Processing Systems*. 2260–2268.

[29] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2390–2395.

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

[31] 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.

[32] 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.

[33] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of naive Bayes text classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 1 (2011), 2.

[34] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernadino, et al. 2019. Learning to assess the quality of stroke rehabilitation exercises. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 218–228.

[35] 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.

[36] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. An exploratory study on techniques for quantitative assessment of stroke rehabilitation exercises. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 303–307.

[37] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Interactive hybrid approach to combine machine and human intelligence for personalized rehabilitation assessment. In *Proceedings of the ACM Conference on Health, Inference, and Learning*. 160–169.

[38] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[39] Andrew F Long, Rosie Kneafsey, and Julia Ryan. 2003. Rehabilitation practice: challenges to effective team working. *International journal of nursing studies* 40, 6 (2003), 663–673.

[40] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[41] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[42] Gail Mountain, Steven Wilson, Christopher Eccleston, Susan Mawson, Jackie Hammerton, Tricia Ware, Huiru Zheng, Richard Davies, Norman Black, Nigel Harris, et al. 2010. Developing and testing a telerehabilitation system for people following stroke: issues of usability. *Journal of Engineering Design* 21, 2-3 (2010), 223–236.

[43] Margit Alt Murphy, Carin Willén, and Katharina S Sunnerhagen. 2011. Kinematic variables quantifying upper-extremity performance after stroke during reaching and drinking from a glass. *Neurorehabilitation and neural repair* 25, 1 (2011), 71–80.

[44] Mark A Musen, Blackford Middleton, and Robert A Greenes. 2014. Clinical decision-support systems. In *Biomedical informatics*. Springer, 643–674.

[45] Susan B O'Sullivan, Thomas J Schmitz, and George Fulk. 2019. *Physical rehabilitation*. FA Davis.

[46] Madhuri Panwar, Dwaipayan Biswas, Harsh Bajaj, Michael Jöbges, Ruth Turk, Koushik Maharatna, and Amit Acharyya. 2019. Rehab-Net: Deep Learning Framework for Arm Movement Classification Using Wearable Sensors for Stroke Rehabilitation. *IEEE Transactions on Biomedical Engineering* 66, 11 (2019), 3026–3037.

[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[48] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[49] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. 2018. Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–16.

[50] Samuele Lo Piano. 2020. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications* 7, 1 (2020), 1–7.

[51] Brandon Rohrer, Susan Fasoli, Hermano Igo Krebs, Richard Hughes, Bruce Volpe, Walter R Frontera, Joel Stein, and Neville Hogan. 2002. Movement smoothness changes during stroke recovery. *Journal of Neuroscience* 22, 18 (2002), 8297–8304.

[52] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy Human-Centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–31.

[53] 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.

[54] Jiliang Tang, Salem Alelyani, and Huan Liu. 2014. Feature selection for classification: A review. *Data classification: Algorithms and applications* (2014), 37.

[55] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25, 1 (2019), 44–56.

[56] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234.

[57] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*.

[58] M Mitchell Waldrop. 2015. Autonomous vehicles: No drivers required. *Nature News* 518, 7537 (2015), 20.

[59] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[60] Elizabeth C Ward, Clare L Burns, Deborah G Theodoros, and Trevor G Russell. 2014. Impact of dysphagia severity on clinical decision making via telerehabilitation. *Telemedicine and e-Health* 20, 4 (2014), 296–303.

[61] David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. 2019. Clinical applications of machine learning algorithms: beyond the black box. *Bmj* 364 (2019).

[62] David Webster and Ozkan Celik. 2014. Systematic review of Kinect applications in elderly care and stroke rehabilitation. *Journal of neuroengineering and rehabilitation* 11, 1 (2014), 108.

[63] Ching-yi Wu, Catherine A Trombly, Keh-chung Lin, and Linda Tickle-Degnen. 2000. A kinematic study of contextual effects on reaching performance in persons with and without stroke: influences of object availability. *Archives of Physical Medicine and Rehabilitation* 81, 1 (2000), 95–101.

[64] 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.

A APPENDIX

Table 2: Parameters of Machine Learning Models

	ROM	Smoothness	Compensation
E1	DT : Depth = 3 LR : Lasso SVM: RBF NN: - Hidden Layers/Units: (32, 32, 32) - Learning Rate: 0.1	DT: Depth 5 LR: Ridge SVM: Linear NN: - Hidden Layers/Units: (16) - Learning Rate: 0.0001	DT: Depth = 5 LR: Ridge SVM: Linear NN: - Hidden Layers/Units: (256, 256) - Learning Rate: 0.1
E2	DT: Depth = 5 LR: Ridge SVM: Linear NN: - Hidden Layers/Units: (256) - Learning Rate: 0.1	DT: Depth = 4 LR: Ridge SVM: Linear NN: - Hidden Layers/Units: (64, 64) - Learning Rate: 0.001	DT: Depth = 3 LR: Ridge SVM: Linear NN: - Hidden Layers/Units: (128, 128) - Learning Rate: 0.1
E3	DT: Depth = 4 LR: Ridge SVM: Linear NN: - Hidden Layers/Units: (256) - Learning Rate: 0.1	DT: Depth = 4 LR: Ridge SVM: Linear NN: - Hidden Layers/Units: (64, 64) - Learning Rate: 0.001	DT: Depth = 3 LR: Ridge SVM: Linear NN: - Hidden Layers/Units: (128, 128) - Learning Rate: 0.1

Table 3: Statistical Analysis on the Results of the User Study

	Responses						Agreement
	Usefulness	Richness	Trust	Effort	Frust	UsageIntent	
Anova Tests	**p < 0.05**	**p < 0.05**	p = 0.19382	**p < 0.05**	p = 0.1962	**p < 0.10**	**p < 0.05**
Traditional vs PredScore	p = 0.1275	p = 0.3576	n/a	p = 0.1290	n/a	p = 0.2154	**p < 0.10**
Traditional vs Proposed	p = 0.1514	**p < 0.10**	n/a	**p < 0.10**	n/a	**p < 0.05**	**p < 0.01**
PredScore vs Proposed	**p < 0.05**	**p < 0.10**	n/a	p = 0.2714	n/a	**p < 0.10**	p = 0.1874