Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment

MIN HUN LEE, Carnegie Mellon University, USA
DANIEL P. SIEWIOREK, Carnegie Mellon University, USA
ASIM SMAILAGIC, Carnegie Mellon University, USA
ALEXANDRE BERNARDINO, Instituto Superior Técnico, Portugal
SERGI BERMÚDEZ I BADIA, University of Madeira, Madeira-ITI, NOVA-LINCS, Portugal

Clinical decision support systems have the potential to improve work flows of experts in practice (e.g. therapist's evidence-based rehabilitation assessment). However, the adoption of these systems is challenging, and the gains of these systems have not fully demonstrated yet. In this paper, we identified the needs of therapists to assess patient's functional abilities (e.g. alternative perspectives with quantitative information on patient's exercise motions). As a result, we co-designed and developed an intelligent decision support system that automatically identifies salient features of assessment using reinforcement learning to assess the quality of motion and generate patient-specific analysis. We evaluated this system with seven therapists using the dataset from 15 patients performing three exercises. The results show that therapists have higher usage intent on our system than a traditional system without patient-specific analysis ($p < 0.05$). While presenting richer information ($p < 0.10$), our system significantly reduces therapists' effort on assessment ($p < 0.10$) and improves their agreement on assessment from 0.66 to 0.71 F1-scores ($p < 0.01$). This work discusses the importance of human centered design and development of a machine learning-based decision support system that presents contextually relevant information and salient explanations on its prediction for better adoption in practice.

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

Additional Key Words and Phrases: Human-AI Interaction; Explainable AI; Machine Learning; Decision Support Systems; Stroke Rehabilitation Assessment;

ACM Reference Format:
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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 156 (October 2020), 27 pages. https://doi.org/10.1145/3415227

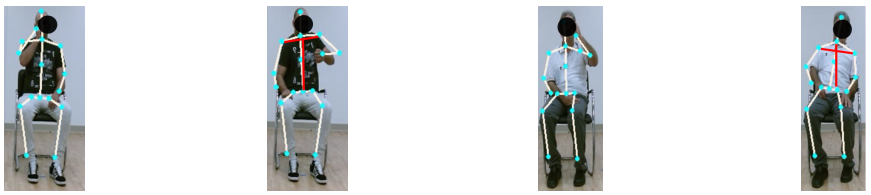
Authors' addresses: Min Hun Lee, minhunl@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Daniel P. Siewiorek, dps@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Asim Smailagic, asim@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Alexandre Bernardino, alex@isr.tecnico.ulisboa.pt, Instituto Superior Técnico, Lisbon, Portugal; Sergi Bermúdez i Badia, sergi.bermudez@m-iti.org, University of Madeira, Madeira-ITI, NOVA-LINCS, Funchal, Portugal.

2573-0142/2020/10-ART156 $15.00
https://doi.org/10.1145/3415227

1 INTRODUCTION

Machine learning algorithms are increasingly being explored and developed in the form of a decision support system to improve various medical domains [7]. One potential application is to support a therapist's evidence-based decision making on assessing the status of a patient with musculoskeletal and neurological diseases (e.g. stroke) [34, 63]. In current practices, therapists typically rely on clinical tests that require their direct, visual observation of patient's exercise motions to evaluate the status of a patient and determine interventions [45]. Although assessment of patient's exercise performance is critical to determine an appropriate intervention, it is infrequently performed due to the limited availability of therapists [40]. Therapists primarily rely on a patient's self-report [45], and have a lack of quantitative data on patient's performance and progress to make informed decision [5, 22]. To address this problem, researchers have demonstrated the potential of technology-assisted decision support tools for rehabilitation, which can monitor patient's exercise motion and assess patient's performance using sensors and machine learning algorithms to generate quantitative measurements for therapist's assessment in a laboratory setting [34, 35, 63].

Despite the potential of clinical decision support systems, prior work has described the challenges of deploying these systems due to the lack of user-centered design consideration [9, 27, 65] and the opaqueness of machine learning algorithms [7, 27, 36, 62]. However, significant recent research has focused on improving the accuracy of monitoring an exercise and replicating clinician's decision making with a complex algorithm [35, 46, 49]. Even if prior work using a complex algorithm makes a system more accurate, it is still difficult to derive a system that can perfectly replicate a therapist's decision making on rehabilitation assessment given diverse physical characteristics of patients [35]. For example, two patients could have different ways of incorrectly performing an exercise (Figure 1). Thus, a system can incorrectly predict a new patient's exercise motion with compensated joints that is not present in the dataset. If a system with complex algorithms cannot provide explanations on its prediction [18, 33], therapists could lose trust in the system and abandon its usage even if it provides valuable predictions in other cases [8, 27, 29].



(a) (b) (c) (d)

Fig. 1. Two patients performing the *'Bring a Hand to Mouth'* exercise with different compensated motions: (a) unaffected and (b) affected motions of a patient 11 (elevated shoulder and trunk rotation). (c) unaffected and (d) affected motions of patient 14 (elevated shoulder and leaning backward).

.

In this paper, we co-design, develop, and evaluate an intelligent decision support system for stroke rehabilitation assessment with therapists (Figure 2). After conducting the interviews and focus group sessions with therapists, we found their needs on rehabilitation assessment, alternative perspectives with quantitative measurements and designed a system accordingly. Given a new patient's affected motion, this system automatically identifies salient kinematic features of assessment to predict the quality of motion and generate patient-specific analysis on a web-based visualization interface (Figure 3). This patient-specific analysis is composed of the predicted quality of motion on three performance components (i.e. *'Range of Motion'*, *'Smoothness'*, and *'Compensation'*), feature analysis

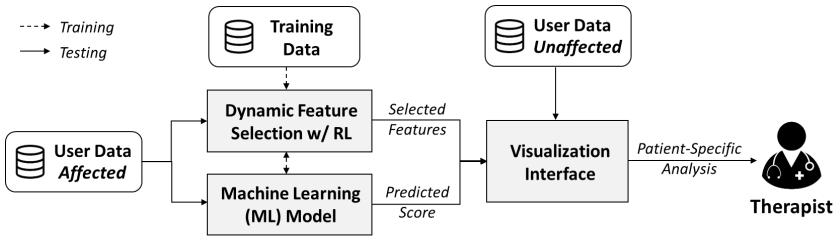


Fig. 2. Flow diagram of an intelligent decision support system: this system automatically selects salient kinematic features of assessment to predict the quality of motion on patient's rehabilitation exercises and generate patient-specific analysis on the web-based visualization interface for therapist's rehabilitation assessment

with identified kinematic features (e.g. joint angle, the trajectory of wrist to the target position, etc. as shown in the Figure 3b), images of salient frames (Figure 3c), and graphs of joint trajectories (Figure 3d) to empower therapist's better understanding and assessment on patient's performance.

For the development of our system, we utilized the dataset of three stroke rehabilitation exercises from 15 post-stroke patients and 11 healthy participants, which were annotated by two therapists. Using this dataset, we applied reinforcement learning to identify the most salient kinematic features of assessment [37] and learn a machine learning model to predict the quality of motion on patient's exercises with leave-one-patient-out cross-validation. After the development, we performed a user study with seven therapists from four rehabilitation centers. Specifically, we investigated how therapists use this system and its effect on therapist's decision making to assess exercise performance of a patient. Results show that therapists had higher usage intent on our system than a traditional system without any user-specific analysis ($p < 0.05$). Our system enabled therapists to validate their assessment with quantitative, user-specific analysis, which increases therapists' utility of our system with richer information ($p < 0.10$) and decreases their effort on an assessment task ($p < 0.10$). In addition, our system assisted therapists to achieve significantly higher agreement on their assessment (0.71 average F1-scores) than the traditional system (0.66 average F1-scores) ($p < 0.01$).

This paper makes the following contributions:

- enumerate needs of therapists during assessing rehabilitation exercises
- describe the co-design and implementation of an intelligent decision support system for stroke rehabilitation assessment that automatically identifies salient features using reinforcement learning to predict the quality of motion and generate patient-specific analysis
- present the quantitative and qualitative evaluation of our system with seven therapists and pose this system as an assistant of therapists to support more consistent assessment

In the following sections, we outline related work and background on the issues of deploying clinical decision support systems, current practices of physical rehabilitation, approaches of technology-assisted rehabilitation systems (Section 2). We then present the study designs for stroke rehabilitation assessment, which includes the findings of therapists' needs and specification on stroke rehabilitation (Section 3), and describe the development of our system (Section 4). Finally, we report the experiments and results of system implementation (Section 5 and 6) and user study (Section 7 and 8), and conclude with discussion on the importance of a human-centered and explainable decision support system for its better adoption in practice.

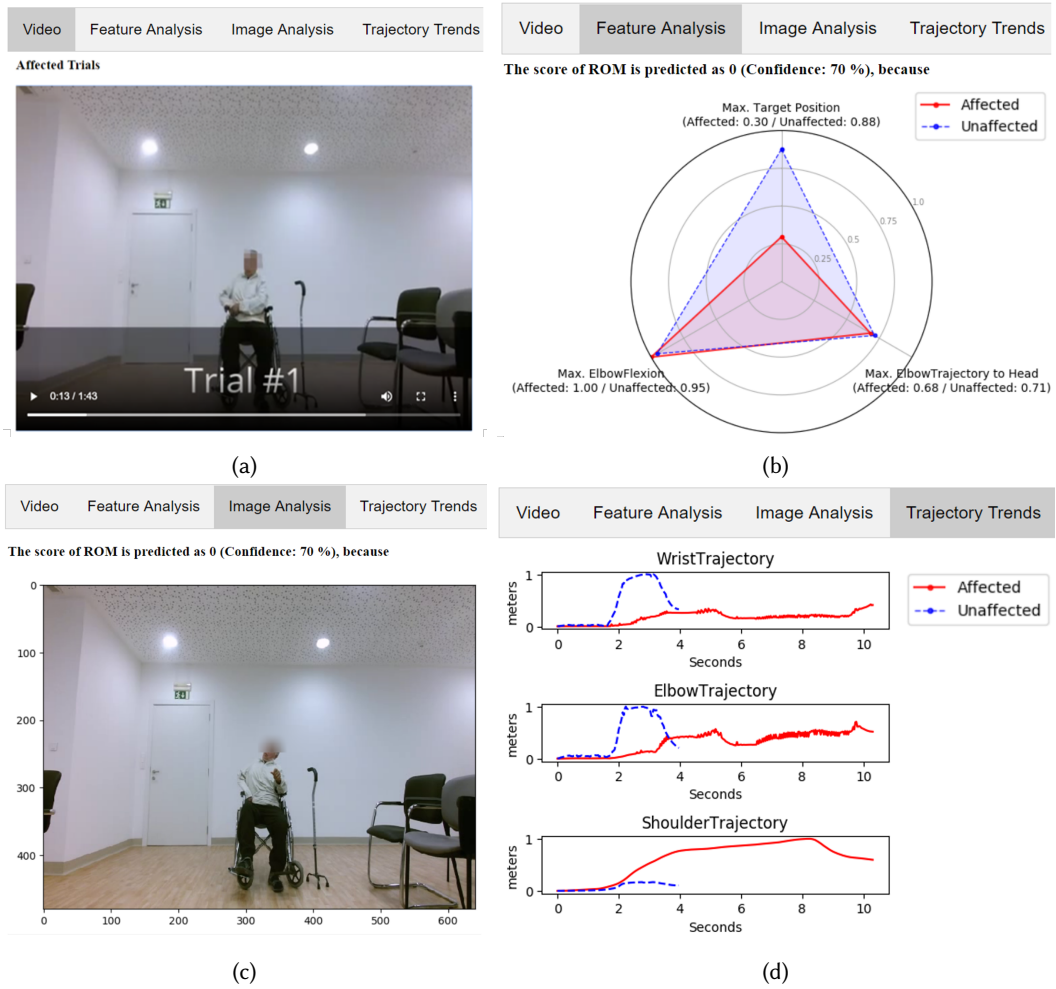


Fig. 3. The web-based visualization interface of the proposed system presents (a) a video of patient's exercise motions and a patient-specific analysis that includes a predicted performance score with (b) feature analysis between unaffected and affected sides, (c) images of salient frames, and (d) graphs of joint trajectories between unaffected and affected sides.

2 RELATED WORK AND BACKGROUND

2.1 Challenges of Deploying Clinical Decision Support Systems

Clinical decision support systems are computational systems that intelligently processes and presents patient-specific information to clinicians, patients, or other individuals to enhance practices of healthcare [44]. These systems can support various tasks, such as generating alerts and reminders for clinicians and patients, selecting treatments, and providing information on medical decision making [16]. Even if such systems have the potential to improve the quality and efficiency of health care [44], there have been several impediments to the adoption of these systems [27] and a lack of studies on the effect of these systems in practice [16].

The major impediments of adopting clinical decision support systems include the lack of user-centered designs [9, 27, 65] and the opaqueness of machine learning algorithms [7, 8, 27, 36, 62]. Clinicians might abandon a system if it does not provide relevant information [65]. Thus, it is recommended to involve users during a development process [26, 27]. When a machine learning algorithm is applied, it is imperative to generate additional explanations and permit clinician's post-hoc analysis on a system [7]. In this work, we focus on a decision support system that augments clinician's decision making. Specifically, we co-design a machine learning-based decision support system for stroke rehabilitation assessment with therapists and study its effect on therapist's rehabilitation assessment.

2.2 Physical Rehabilitation for Post-Stroke Patients

Patients with musculoskeletal and neurological disorders (e.g. stroke) require a rehabilitation program over several months to improve their functional abilities [45]. During a rehabilitation program, therapists first diagnose the condition of a patient with various methods (e.g. analyzing patient's history, conducting tests, or analyzing measurements). Therapists usually rely on clinical tests that leverage their direct, visual observation or video recordings of motions from a patient to evaluate motor symptoms [17, 25, 56, 59]. Based on a test with direct visual observation, therapists identify dysfunctional areas of patient's body and determine interventions [45]. In case of post-stroke patients, therapists prescribe to perform a task-oriented exercise (e.g. bring a cup to mouth), which is one of the effective interventions [50]. After interventions, therapists discuss patient's progress to modify interventions as appropriate [45].

Although assessing patient's performance on rehabilitation exercises is important for therapists to adjust interventions, this assessment relies on therapist's experience [53] and infrequently performed due to the limited availability of a therapist [40]. Therapists primarily reply on a patient's self-report or brief visual observation to understand how well a patient follows the prescribed interventions [45]. They have a lack of quantitative data on patient's performance and progress [22]. Thus, therapists encounter challenges of understanding the patient's performance and making informed decisions to adjust intervention [5].

2.3 Technological Support for Physical Rehabilitation

Researchers have investigated various technologies to facilitate the delivery of physical rehabilitation [41]. In the following sections, we outline existing motion tracking techniques, applications for patients and therapists, and motion analysis techniques for rehabilitation monitoring and assessment.

2.3.1 Motion Tracking Techniques.

One primary technology of rehabilitation is a motion tracking system that dynamically represents the pose of a human body using sensors. Motion tracking systems have been developed with various techniques: non-visual sensors (e.g. inertial, magnetic, etc.) and visual marker-based or marker-free [66]. A visual marker-based technique utilizes infrared cameras capturing motions from the reflective markers on the human body. Among various systems, a visual marker-based system has the highest performance (errors are around 1mm) and is often used as a golden standard due to their accurate position information [66]. Some specialized centers and clinics have adopted motion analysis tools based on this visual marker-based system (e.g. VICON or Optitrack) [57]. However, even if this visual marker-based system provides high-precision data, they have a complex set-up that requires expert operation and expensive costs [11]. In contrast to visual marker-based systems, both non-visual, inertia sensors and visual marker-free system provide competitive tracking performance for rehabilitation monitoring [10, 11, 63, 66] and lower cost for patients and clinicians

[10, 11, 63]. As inertia sensors have limitations of measurement noise due to inconsistent positions of sensors and cumbersome to wear sensors, this work applies a visual marker-free technique (i.e. a Kinect sensor) to track patient's rehabilitation exercises.

2.3.2 *Applications for Patients and Therapists.*

Motion tracking techniques can be realized into different applications for two major stakeholders in rehabilitation: patients and therapists. For better patient's rehabilitation experiences, researchers have explored alternative interventions with exoskeleton, socially assistive robots [13, 39] and virtual reality [43]. Building upon a motion tracking technique with sensors, exoskeleton robots apply mechanical support to the patient's body to induce patient's passive motion [39], or socially assistive robots can provide corrective feedback on patient's exercise [13, 33, 38]. Virtual reality based rehabilitation systems use computer-simulated interactive environments to promote patient's movement [43]. In addition, with increasing emphasis on evidence-based decision making about patient care and assessment [5], researchers have developed a decision support system for therapists [34]. Given human motion data of a motion tracking technique, decision support systems simply process data to provide quantitative measurements or apply motion analysis techniques to recognize and assess symptoms of a chronic disease [34, 63].

2.3.3 *Motion Analysis Techniques for Rehabilitation Monitoring and Assessment.*

Motion analysis techniques for rehabilitation monitoring and assessment can be categorized into either a rule-based model or a machine learning model [36]. A rule-based model is based on a set of monitoring rules through the involvement of a domain expert [36] . For example, researchers compared the positions of wrist and spine joints to monitor the completion of an upper-limb exercise [36]. This rule-based approach provides the modularization and flexibility to develop a monitoring system. However, it is time consuming [2] to determine the right threshold values of rules for an individual's status. Moreover, experts might not be able to articulate their heuristic-based decision making on a complex monitoring task [2]. Alternative approach utilizes a machine learning algorithm to process complex sensor data and automatically extract a meaningful function (e.g. Neural Network model) that can classify the quality of motion [35, 46, 55]. However, no algorithm can completely replicate a therapist's assessment given diverse physical characteristics and functional abilities of patients. Moreover, a machine learning model with complex algorithms cannot explain its prediction to support expert's decision making [18, 33], which can exacerbate therapist's experience with a decision support system [7, 23, 27, 62]. In this paper, we aim to increase the interpretability of a model by feature selection [6, 28]. Specifically, we apply reinforcement learning [37] to identify kinematic salient features for assessment. Utilizing an identified subset of features, we predict the quality of motion and generate patient-specific analysis to summarize exercise performance of a patient for therapist's assessment.

Although Mansoor et al. discusses the necessity of more investigation on challenges of accepting patient monitoring systems in clinic [3], a significant recent research has focused on improving the accuracy of monitoring and replicating clinician's decision making with a complex algorithm [35, 46, 49]. There is the lack of knowledge and evaluation [16] about therapist's experience on a decision support system for physical rehabilitation monitoring and assessment. In this work, we co-design an intelligent decision support system for stroke rehabilitation assessment with therapists to accommodate their needs and provide relevant information. In addition, we study the effect of the system on therapist's rehabilitation assessment. This study contributes to increase knowledge on the effects and values of a human-centered, intelligent decision support system in practice.

3 STROKE REHABILITATION AS A TEST DOMAIN

We selected a probe domain as stroke, which is the second leading cause of death and third most common contributor to disability [14]. We recruited nine therapists of stroke rehabilitation from five rehabilitation centers (Table 1), and iteratively involved them in different phases of our study (i.e. designs, implementation, and evaluation). In this section, we first present the designs of our study through two activities: 1) understanding therapists' needs during stroke rehabilitation assessment through interviews and focus group sessions with nine therapists and 2) specification of a study with three therapists (i.e. exercises and performance components for assessment). The other involvements of therapists (i.e. annotation, review and evaluation on system implementation) will be described in Section 5 and 7.

Table 1. Profiles of Therapists in Various Activities of the Study: need finding (needs), specification, annotation, review on the interface (review), and evaluation

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

3.1 Needs during Rehabilitation Assessment

We observed an one hour-long rehabilitation session at one rehabilitation center and performed 3 interviews and 2 focus groups (group 1: $n = 2$; and group 2: $n = 4$) with nine therapists from five rehabilitation centers to gain knowledge about the current practices and therapists' needs of assessing patient's rehabilitation exercises. Therapists were recruited by email communications to local hospitals through advertising to university staff and contacts of the research team. To collect diverse opinions during an interview or a focus group session, we recruited nine therapists (2 males and 7 females; 29.6 ± 5.4 years old) with various experiences and disciplines: $\mu = 7.44$, $\sigma = 5.51$ years in stroke rehabilitation; 6 occupational therapists, who focus on helping patients better engagement in daily life and 3 physiotherapist, who treats patient's actual impairment from a bio-mechanical perspective as shown in Table 1.

An experimenter (the first author) moderated a semi-structured interview or focus group session for an hour with the following topics: the process of assessment (*"what's your process of monitoring patient's functional ability in practice?"*), strategies to cope with an uncertain situation (*"what do you do if you are uncertain?"*), the current usage of technology (*"any technical tool is being used?"*), and opportunities of technical support (*"do you have any information or support that you would you like to receive from a system?"*). For the data analysis, we transcribed interviews and focus group sessions and followed an iterative coding process [15]. We first created initial codes from literature review and questions. Two researchers then independently coded a subset of the transcript data with initial codes and generated additional codes if required. Then, the results were discussed to generate the final codes for the data analysis. The thematic analysis on need findings with therapists includes 1) their challenges with time-consuming, infrequent, and experience-based

assessment and 2) their desires for alternative perspectives with quantitative measurements through simple and intuitive technologies.

3.1.1 Time-consuming, Infrequent, and Experience-based Assessment.

For the management of physical rehabilitation, therapists rely on an iterative process that examines the patient's status, determines an intervention (e.g. a set of exercises), and reexamine the treatment outcomes and the patient's status for adjustment [45]. The examination on the patient's status is only conducted every two or three months as patient's progress is typically slow and the examination takes a long time [45]. *"I do not perform the examination at every session. Instead, we typically do every three months unless there is any occurrence of a special issue or situation"* (TP 3). After determining an intervention, a therapist arranges a weekly session with a patient. During a session, therapists mainly rely on their observation and experience to approximately assess the performance of a patient and provide feedback.

3.1.2 Challenges of Making Informed Decision Making.

During observation and experience-based assessment, therapists encountered various challenges to make informed decision making. Therapists commented that *"there is no exact single definition of normality for assessment"* (TP 1). Instead, therapists mentioned that they typically *"check the functionality of the unaffected side and define adequate normality for each patient"* (TP 9). After internally generating hypothetical correctness of a movement with the patient's unaffected side, therapists then *"analyze various aspects of performance: whether a patient can complete an expected movement and any compensated, not coordinated movement exists"* (TP 2). For instance, according to our observation on a rehabilitation session, a therapist first asked a patient to perform a motion multiple times or keep at a certain position for a while for assessment. A therapist then had to keep moving front, back, and side to collect evidence for assessment. Therapists expressed a *"difficulty with collecting information on patient's rehabilitation exercise performance"* (TP 3). In addition, therapists have uncertainty to evaluate an abstract performance component (i.e. smoothness). TP 1 commented that therapists have *"difficulty with assessing smoothness of a motion"* as the smoothness of a motion can be differently interpreted and defined by each therapist. When therapists are unsure, they mentioned that they record patient's movements to review and *"re-evaluate more confidently after a session by watching a video multiple times"* (TP 7), or *"discuss with other colleagues"* (TP 8) on their experience-based assessment.

3.1.3 Desire for Alternative Perspectives with Quantitative Measurements.

When discussing opportunities of technological supports to assess rehabilitation exercises, therapists referred to the need to gain insights on patient's performance with *"alternative perspectives on assessment and quantitative kinematic measurements"* (TP 6). Instead of relying only on the therapist's own assessment, therapists commented whether a system can provide *"alternative assessment"* (TP 8) for validation. In addition, therapists need quantitative measurements to *"detect minor changes over time and discuss with other colleagues"* (TP 2) on an performance component. Specifically, therapists want to know *"how closely a patient can reach a target motion"*, *"how smoothly motion is coordinated"*, and *"to which extent a patient performs a compensated motion"* (e.g. *"how much a shoulder joint is elevated"*) (TP 3) with quantitative measurements and images of a patient's motion. In addition, TP 1 commented that *"trajectory trends (e.g. showing a graph on how a wrist joint trajectory changes during a motion) would be useful to understand smoothness of motion"*.

3.1.4 Considerations of Technological Support: Simple and Intuitive Usage.

Although therapists mostly expressed the potential benefit of technical support on rehabilitation assessment, all rehabilitation centers that we visited or discussed do not use any technology for managing stroke rehabilitation. Despite of the low adoption of technology in practice, therapists

commented that there were previous attempts to use technology for rehabilitation assessment. *"We tried to use few technologies for rehabilitation before"*. However, these technologies (e.g. a posturography system [61]) are *"expensive"* (TP 3) and *"require many complex and time-consuming steps for the usage. As we have limited time to interact with patients during a session, we do not want to waste time on setting up a system. We ended up discarding the usage of a complex system in practice." (TP 9)*. TP 9 emphasized that *"a system should be easy to use and present insights quickly with intuitive graphics given the limited session time for each patient."*, so that it can fit within a time-constrained session in practice.

3.1.5 Requirements of a System.

Based on our need findings with therapists, we identified the design requirements of an intelligent decision support system for stroke rehabilitation in Table 2. First, a system should provide comparison between unaffected and affected sides of a patient so that it can facilitate therapists to ground normality of an individual patient for assessment. In addition, a system should provide not only predicted assessment, but also quantitative measurements to support the therapist's informed decision making on assessment. Lastly, a system should provide simple and intuitive graphics so that therapists can utilize it quickly in practice. The details of our implementation are described in Section 4.3.

Table 2. The list of needs from therapists and corresponding requirements of an intelligent decision support system

Needs of Therapists	Requirements of a System
N1. Define normality with unaffected motions of a patient	R1. Comparison between unaffected and affected motions
N2. Validate assessment with another perspective of assessment	R2. Predicted assessment from a model calibrated with another therapist
N3. Collect information on patient's performance - N3.1. Detect minor changes - N3.2. Watch a video multiple times - N3.3. Understand smoothness of a motion	R3. Present additional patient-specific analysis - R3.1. Quantitative kinematic measurements - R3.2. Images of a patient's motion - R3.3. Graphs of joint trajectories
N4. Simple, intuitive usage and presentation	R4. Avoid overwhelming and utilize graphics to present insights quickly

3.2 Specifications

After having iterative discussion with three therapists (TPs with check marks in the *'Specification'* column of Table 1; $\mu = 6.49$, $\sigma = 2.05$ years of experience in stroke rehabilitation), we specified exercises and performance components of assessment to probe how therapists utilize an intelligent decision support system to assess patient's rehabilitation exercises.

3.2.1 Three Task-Oriented Upper Limb Exercises.

This work utilizes three upper-limb stroke rehabilitation exercises (Figure 4), recommended by therapists. In Figure 4, the *'Initial'* label indicates the initial position of an exercise and the *'Target'* label describes the desired task position of an exercise.

For Exercise 1, a patient has to raise his/her wrist to the mouth as if drinking water. For Exercise 2, a patient has to pretend to touch a light switch on the wall. Exercise 3 is to practice the usage of a cane while extending the elbow in the seated position. These exercises are selected due to their correspondence with major motion patterns: elbow flexion for Exercise 1, shoulder flexion for Exercise 2, elbow extension for Exercise 3.

3.2.2 Performance Components.

After reviewing popular stroke assessment tools (i.e. Fugl Meyer Assessment [56] and Wolf Motor Function Test [59]) and having iterative discussions with therapists, we identified three common

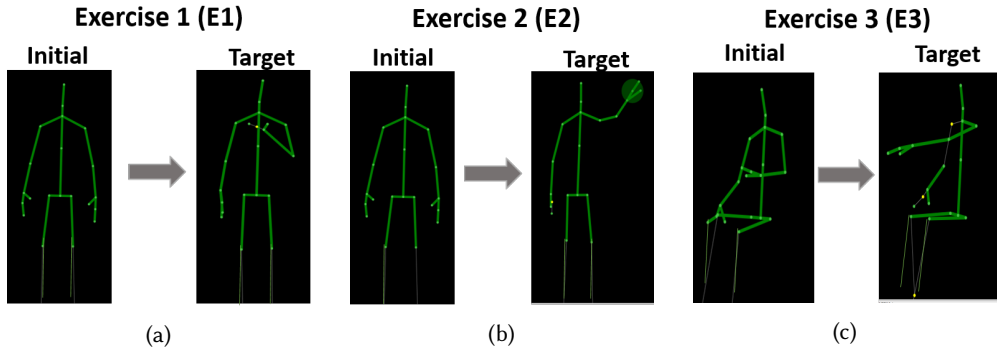


Fig. 4. (a) Exercise 1 (E1): *'Bring a Cup to the Mouth'* (b) Exercise 2 (E2): *'Switch a Light On'* (c) Exercise 3 (E3): *'Move a Cane Forward'*

performance components and their scoring guidelines: *'Range of Motion (ROM)'*, *'Smoothness'*, and *'Compensation'* (Table 3). The *'ROM'* component describes the amount of a joint movement, how closely a patient achieve a task-oriented exercise. The *'Smoothness'* component indicates the degree of a trembling and irregular movement of joints while performing an exercise. The *'Compensation'* component checks whether unnecessary joints are used to achieve a target movement. For instance, a patient might elevate the shoulder, lean backward, or rotate the trunk to raise the affected hand as shown in Figure 1b and 1d.

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
Smoothness	0	Excessive tremor or not smooth coordination
	1	Movement influenced by tremor
	2	Smoothly coordinated movement
Compensation	0	Noticeable compensation in more than two joints
	1	Noticeable compensation in a joint
	2	Does not involve any compensations

3.2.3 Kinematic Features.

We represent patient's exercises with sequential joint coordinates from a Kinect v2 sensor (Microsoft, Redmond, USA) and extract various kinematic features based on literature review [35, 51, 56, 59] and discussions with therapists.

To represent the *'ROM'* component, we extract joint angles (e.g. elbow flexion, shoulder flexion, elbow extension), normalized relative trajectory (i.e. Euclidean distance between two joints - head and wrist, head and elbow), and normalized trajectory distance (i.e. absolute distance between two joints - head and wrist, shoulder and wrist) in the x, y, z coordinates [35].

For the *'Smoothness'* component, we compute various speed related features: speed, acceleration, and jerk, zero crossing ratio of acceleration and jerk, and Mean Arrest Period Ratio (i.e. the portion

of the frames when speed exceeds 10% of the maximum speed) [35, 51]. As this study mainly utilizes upper-limb exercises, we computed these speed related features on wrist and elbow joints.

For the *'Compensation'* component, we compute joint angles (i.e. the elevated angle of a shoulder, the tilted angle of spine, and shoulder abduction) and normalized trajectories (distances between joint positions of head, spine, shoulder in the x, y, z coordinates from the initial to current frames) to distinguish the occurrence of a compensated movement [35, 56, 59].

Before extracting features, we apply a moving average filter with the window size of five frames to reduce noise of acquiring joint positions from a Kinect sensor similar to [35]. For each exercise motion, we compute a feature matrix ($\mathbf{F} \in R^{t \times d}$) with t frame and d features and 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 INTELLIGENT DECISION SUPPORT SYSTEM FOR STROKE REHABILITATION ASSESSMENT

Based on identified therapists' needs, we designed and implemented an intelligent decision support system (Figure 2) that identifies salient features for assessment using reinforcement learning to predict the quality of motion and generate user-specific analysis that includes feature analysis, images of salient frames, and graphs of joint trajectories (Figure 3). This system enables therapists to review alternative perspectives on patient's exercise performance with quantitative user-specific analysis for assessment.

4.1 Feature Selection using Reinforcement Learning

Kinematic variables analysis is an important way for therapists to quantitatively understand the performance of a patient [64]. Yet, simply presenting all variables can overwhelm therapists and limit their ability to gain insights on the performance of a patient. Given the limited availability to administrate multiple patients, therapists want to minimize the amount of time on analyzing kinematic variables while accurately diagnosing the status of a patient. Thus, we aim to automatically identify salient features of assessment with a machine learning technique.

The classical approaches of feature selection (e.g. filter, wrapper, embedded methods) [58] find a fixed feature set to the entire dataset, which applies uniformly to all patients. Instead, this paper utilizes a Markov Decision Process (MDP) to dynamically select a feature set for each patient's motions. As each patient has different physical and functional status (Figure 1), we hypothesize that feature selection with MDP can be beneficial over classical feature selection approaches for personalized rehabilitation assessment.

4.1.1 *Problem Definition.*

We mathematically describe the Markov Decision Process (MDP) with similar notations of [24, 37] as follows:

Let $(X, Y) \in \mathcal{D} = \mathcal{X} \times \mathcal{Y}$ be a sample from a dataset, where X is a feature vector and 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 an observed state of the agent, recruited feature without the label be $s' = \{(x_i, f_i) \mid \forall\, f_i \in \mathcal{F}\}$

- **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 $\lambda = 0.01$. In addition, the agent receives a reward of -1 for incorrect classification and a reward of 0 for correct classification.
- **Transition**: Let the transition function be
$$p(s,a) = \begin{cases} (X, Y, \bar{\mathcal{F}} \cup f_i) & \text{if } a \in \mathcal{A}_f \text{ and } a = f_i \\ TS & \text{if } a \in \mathcal{A}_c \end{cases},$$
where TS is the terminal state after outputting the classification and revealing the true label.

Each episode of this MDP is to classify an instance 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. To solve this problem, we apply Q-network with Double Q-learning [60] using *'PyTorch'* libraries [47]. The architecture and parameters of a neural network for Q-learning are described in Table 6. The implementation details can be found in Appendix A.1

4.2 Machine Learning Model

A machine learning (ML) model applies a supervised learning algorithm to predict the quality of motion on each performance component. We explore various supervised learning algorithms: Decision Trees (DTs), Linear Regression (LR), Support Vector Machine (SVM) using the *'Scikit-learn'* [48] library and Neural Networks (NNs) using *'PyTorch'* [47] library.

For DTs, we implement Classification and Regression Trees (CART) to build prune trees while grid-searching different the maximum depth size of a tree (i.e. 3 - 5). For LR models, we apply $L1, L2$ regularization or linear combination of $L1$ and $L2$ (ElasticNet with 0.5 ratio) to avoid over-fitting. For SVMs, we apply either linear or Radial Basis Function (RBF) kernels with penalty parameters, $C = 1.0$. For NNs, we grid-search 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. $0.0001, 0.005, 0.001, 0.01, 0.1$). We apply *'ReLu'* activation functions and *'AdamOptimizer'* and train a model until the tolerance of optimization is 0.0001 or the maximum 200 iterations.

4.3 Visualization Interface

Based on the therapists' needs (Section 3.1) and guidelines of Human Artificial Intelligence (AI) interaction [1, 32], we design and implement the web-based visualization interface that presents a predicted performance score and user-specific analysis that contains feature analysis, images of salient frames, and graphs of joint trajectories (Figure 3) to support therapist's assessment. Feature analysis shows quantitative difference between unaffected and affected sides using identified salient features (Figure 3b). Images of salient frame show patient's motion at salient frames, in which salient features occur (Figure 3c). Graphs of joint trajectories describe overall trends and duration of joint trajectories during a motion (Figure 3d). The interface has the tab menus to present videos and each analysis respectively. We implement the javascript functions to count the video events (e.g. *'Play'*, *'Pause'*) and measure the amount of time that a user spends on each page during assessment.

As therapists desire other perspectives on assessment to validate their own assessment ($N2$ in the Table 2), this interface presents the predicted assessment, scores on performance components. When presenting this predicted performance score, the performance of our system for predictions is also included to *"make clear how well the system can do"* [1]. In addition, this interface presents user-specific analysis that is considered *"contextually relevant information"* [1] on patient's exercise

performance ($N3$ in the Table 2). This user-specific analysis includes the presentation of feature analysis, images of salient frames, and graphs of joint trajectories that are identified during the needs finding study with therapists (Section 3.1). Feature analysis supports therapists to detect minor changes on a patient quantitatively ($N3.1$ in the Table 2). Utilizing selected salient features (e.g. the maximum target position, maximum elbow flexion in Figure 3b), we identify frames in which these salient features occur to present images. These images of salient frames aim at facilitating therapist's search on evidence from a video ($N3.2$ in the Table 2). In addition, as observing sequential patterns of kinematic variables facilitate understanding on the smoothness of a motion ($N3.3$ in the Table 2), this interface shows trajectories of three major joints (e.g. shoulder, elbow, and wrist) for upper-limb exercises. Throughout this paper, we refer the user-specific analysis with identified kinematic features as explanations of predicted assessment.

For simple and intuitive presentation ($N4$ in Table 2) on identified salient features, this interface utilizes a radar chart to effectively present multivariate data [52]. To *"avoid overwhelming"* [32] therapists, this interface limits to include only three salient features with highest information gain. As therapists utilize patient's unaffected motion as normality to assess patient's performance ($N1$ in Table 2), this interface follows this current practice, *"social norms"* [1] and includes the comparison between the affected and unaffected sides of an individual patient to present salient features and graphs of joint trajectories.

5 SYSTEM IMPLEMENTATION STUDY

5.1 Data Collection

We recruited 15 stroke patients and 11 healthy participants to collect the dataset of three upper limb exercises using a Kinect v2 sensor (Microsoft, Redmond, USA). The data collection program was implemented in C# using Kinect SDK and Accord.NET framework and operated on a PC with 8GB RAM and i5-4590 3.3GHz 4 Cores CPU [35]. This program records the 3D trajectory of joints and video frames at 30 Hz. The sensor was located at a height of 0.72m above the floor and 2.5m away from a participant. The starting and ending frames of exercise movements were manually annotated during the data collection.

Before participating in the data collection, all participants signed the consent form. Fifteen post-stroke patients (13 males and 2 females in Table 5) participated in two sessions for data collection: During the first session, a therapist evaluated post-stroke patient's functional ability using the a clinically validated tool, Fugl Meyer Assessment (FMA) (the maximum score on 66 points) [56]. Fifteen post-stroke patients have diverse functional abilities from mild to severe impairment (37 ± 21 Fugl Meyer Scores). During the second session, a stroke patient performed 10 repetitions of each exercise with both affected and unaffected sides. Eleven healthy participants (10 males and 1 female) performed 15 repetitions with their dominant arms for each exercise.

After collecting the dataset, 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 watched only the recorded videos of the patient's exercises (Figure 3a) and annotated the performance components of the patient's exercises using the scoring guideline (Table 3) without reviewing analysis of our system (Figure 3d, 3c, 3d). The annotations of TP 1 and 2 are compared to compute the agreement level of therapists. Therapists had fairly good agreement on annotations (Cohen's kappa [42, 54], $\kappa = 0.69$).

5.2 Evaluation Method

For implementation, we utilized the annotation of therapist 1 (TP 1), who had more interactions with recruited stroke patients by supporting the recruitment and evaluation on their functional

ability with Fugl Meyer Assessment [56]. The dataset of each exercise is divided into *'Training'* and *'User'* data to implement our system.

'Training Data' (Figure 2) 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 2) includes each testing stroke patient's unaffected and affected motions. With patient's affected motions, our approach selects patient-specific features and predicts the quality of motion on performance components. Both unaffected and affected motions of a patient are utilized to generate user-specific analysis of the visualization interface (Figure 3).

We applied leave-one-patient-out cross validation on post-stroke patients to implement and evaluate our feature selection and machine learning models. A model was trained with data from all participants except one post-stroke patient and test 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 computes the harmonic mean of precision and recall: $\frac{2*(recall*precision)}{(recall+precision)}$, where precision indicates $\frac{TruePositive}{TruePositive+FalsePositive}$ and recall describes $\frac{TruePositive}{TruePositive+FalseNegative}$. The F1-score 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.

We conducted paired t-tests over three performance components of three exercises to compare performance of our approach with various machine learning (ML) models. In addition, we indicated the agreement level of therapists as TPA in Table 4 with F1-scores to analyze performances of various ML models. After implementing the system and interface, two therapists (TP 1 and 2 with check marks in the *'Review'* column of Table 1) reviewed the web-based visualization interface to detect any problems and improve its usability.

6 SYSTEM IMPLEMENTATION RESULTS

6.1 Feature Selection and Machine Learning Models

Table 4 summarizes average F1-scores of therapist's agreement (TPA) and various machine learning models: a decision tree (ML-DT), linear regression (ML-LR), and a support vector machine (ML-SVM), and a neural network with reinforcement learning-based feature selection (ML-RL). The parameters of machine learning models that achieve the best F1-scores during leave-one-subject-out cross validation are summarized in the Table 6.

The ML-RL model achieves a decent agreement level with ground truth annotation: 0.8119 average F1-score over three exercises. In addition, the ML-RL model outperforms other algorithms with statistical significance ($p < 0.01$ using the paired t-tests over three exercises and three performance components): ML-DT with 0.7011 average F1-score, ML-LR with 0.6981 average F1-score, and ML-SVM with 0.7204 average F1-score. In addition, compared to a model with Recursive Feature Elimination (ML-RFE) method [19], one of classical feature selection methods, ML-RL has 0.11 higher average F1-score ($p < 0.01$ using the paired t-tests over three exercises and three performance components) and is expected to be more beneficial to generate patient-specific analysis for therapists. Compared to the therapists' agreement (TPA) between therapist 1 and therapist 2, ML-RL has 0.04 higher average F1-score. However, the difference of performances between the TPA and ML-RL is not statistically significant.

6.2 Design Review on the Interface

According to the review on our visualization interface, therapists (TP 1 and 2) had problems with understanding the name of features in feature analysis. Thus, we reviewed the names of features

Table 4. Performance (F1-scores) of machine learning (ML) models and therapist's agreement (TPA) between TP1 and TP2. *** indicates that ML-RL has statistically superior performance than the compared method using the pairwise test at 99% significance level.

	Exercise 1 (E1)	Exercise 2 (E2)	Exercise 3 (E3)	Overall
ML - DT ***	0.6901 ± 0.0405	0.7645 ± 0.0867	0.6488 ± 0.0412	0.7011 ± 0.0769
ML - LR ***	0.7246 ± 0.0593	0.6430 ± 0.0982	0.7267 ± 0.0391	0.6981 ± 0.0801
ML - SVM ***	0.7232 ± 0.0364	0.6971 ± 0.0891	0.7410 ± 0.0052	0.7204 ± 0.0585
ML - RFE ***	0.6742 ± 0.0715	0.7628 ± 0.1708	0.6415 ± 0.0806	0.6928 ± 0.1147
ML - RL	**0.8331 ± 0.0059**	**0.7973 ± 0.0868**	**0.8054 ± 0.0496**	**0.8119 ± 0.0526**
TPA	0.7455 ± 0.2054	0.8147 ± 0.1522	0.7254 ± 0.1838	0.7619 ± 0.1626

with TP 1 and 2 and converted them into clinically relevant and easily comprehensive terminologies. For this conversion of feature names, we presented all feature names and described what each feature measures to TP 1 and 2. They spoke aloud how they would describe each feature. For instance, the feature *'normalized trajectory distance of spine x'* is converted to the term *'leaning trunk to the side'*. This feature can be referred to as *'lateral spine flexion'* with a clinical terminology. However, therapists suggested to refer to this feature with a more comprehensive phrase *'leaning trunk to the side'* as they desired to discuss feature analysis with patients, who are not familiar with clinical terms.

7 REAL-WORLD USER STUDY

We performed a user study to investigate how the information of an intelligent decision support system (e.g. predicted performance scores with feature analysis, images of salient frames, graphs of joint trajectories) affects therapist's rehabilitation assessment. For the user study, we compared the experiences of therapists using our proposed interface (Figure 3) to two baseline interfaces: *'Traditional'* interface that presents only videos for assessment and *'Predicted Scores (PredScore)'* interface that presents videos with predicted scores without any user-specific analysis. Specifically, this study aims to address the following questions:

- How do predicted assessment and user-specific analysis of our interface affect the utility of information for decision making and the attitude on our interface (e.g. trust, workload, usage intention, preference) compared to two baseline interfaces (one with only videos and the other with videos and only predicted assessment)?
- Do predicted assessment and user-specific analysis of our interface support more consistent assessment?

In the study, we referred to interfaces as "condition 1", "condition 2", and "condition 3" (counter-balanced) to avoid biasing participants. In this paper, we refer to them as the "traditional", "predicted score (predscore)", "proposed" interfaces for clarity.

7.1 Metrics

We evaluated three interfaces with respect to the following metrics: 1) responses on questionnaires from therapists, 2) logs of a web-based visualization interface, 3) their agreement level on assessment (F1-scores).

7.1.1 *Questionnaires on an Interface.*

We utilize the following questionnaires [8] to collect therapist's opinions on interfaces. All questionnaires were rated on a 7-point scale.

- Usefulness: *"[Tool - Condition X] is useful to understand and assess patient's performance"*
- Richness: *"[Tool - Condition X] generates new insights on patient's performance"*
- Trust: *"I can trust information from [Tool - 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 [Tool - Condition X]?"* and *"How insecure, discouraged, imitated, stressed were you while using the [Tool - Condition X]?"*
- Usage Intention: *"I would use [Tool - Condition X] to understand and assess patient's performance"*
- Preference between two interfaces: participants rated on a 7-point scale ranging from 1 (totally Condition X), 2 (much more Condition X than Y), 3 (slightly more Condition X than Y), 4 (neutral), ..., 7 (totally Condition Y). The preference is asked pairwise on three interfaces: traditional, predscore, and proposed.

7.1.2 *Logs of an Interface.*

All web-based interfaces record a log that measures the amount of time that a therapist spends on each page/resource during assessment and counts the number of video events (e.g. *'Play'*, *'Pause'*).

7.1.3 *Agreement Level of Therapists' Assessment.*

Even if most medical disciplines rely on standardized guidelines [17, 25, 56, 59], experts can be biased in their decision making based on their own experiences and expert disagreement is prevalent [4, 30, 31]. Thus, this study utilizes the level of agreement of therapists' assessment (F1-score) as a metric to analyze the effect of a decision support system. Therapists generate assessment on patient's exercise performances while using each interface. We utilize this therapists' assessment to compute the agreement level of therapists, and explore whether our proposed interface with user-specific analysis supports therapists more consistent assessment.

7.2 Method

Seven therapists (with $\mu = 8.14$, $\sigma = 6.05$ years of experience in stroke rehabilitation) from four rehabilitation centers participated in the user study (TPs with check marks in the *'Evaluation'* column of Table 1). Note that we excluded two therapists (TP 1 and 2), who generated annotations to implement our system and reviewed the design of the interface. After signing an informed consent (Institutional Review Board approved), each therapist was instructed on the procedure of the study and three interfaces using dummy data (30 minutes). Then, a therapist was assigned to the task of assessing 45 videos (around one minute per video, in which a patient performs a rehabilitation exercise) using three interfaces (1.5 hours total) and followed by post-study questionnaires and interviews (30 minutes).

Each interface is assigned to a sub-task of assessing 15 videos (five patients performing three exercises). Therapists 1, who evaluated functional abilities of 15 patients, divided patients into three sub-groups, in which patients of each subgroup have similar functional ability. Thus, the sub-task of each interface is counterbalanced. The order of the three conditions/interfaces and assignment of a sub-task are randomized. After completing a sub-task on each interface, therapists responded to the questionnaires. After finishing all sub-tasks, therapists answered the preference questionnaires. Post-interviews were conducted to understand therapists' perspectives on the effectiveness of the proposed, intelligent decision support system and opportunities to utilize this system in practice.

8 USER STUDY RESULTS

For analysis of results, we first performed one-way ANOVA tests on results of metrics (i.e. responses on questionnaires from therapists, logs of the web interfaces, and their agreement level on assessment). If the results have any statistical significance, we further performed pairwise statistical analysis on three conditions/interfaces 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 7.

8.1 Responses on Questionnaires

Figure 5 summarizes the responses of questionnaires from therapists on three interfaces: "traditional", "predicted scores (predscore)", and "proposed" interfaces. We indicated any statistical significance using one-way ANOVA tests at the bottom of each variable. If a variable has any statistical significance using ANOVA tests, we also denoted any statistical significance using post-hoc paired t-tests at the top of each variable.

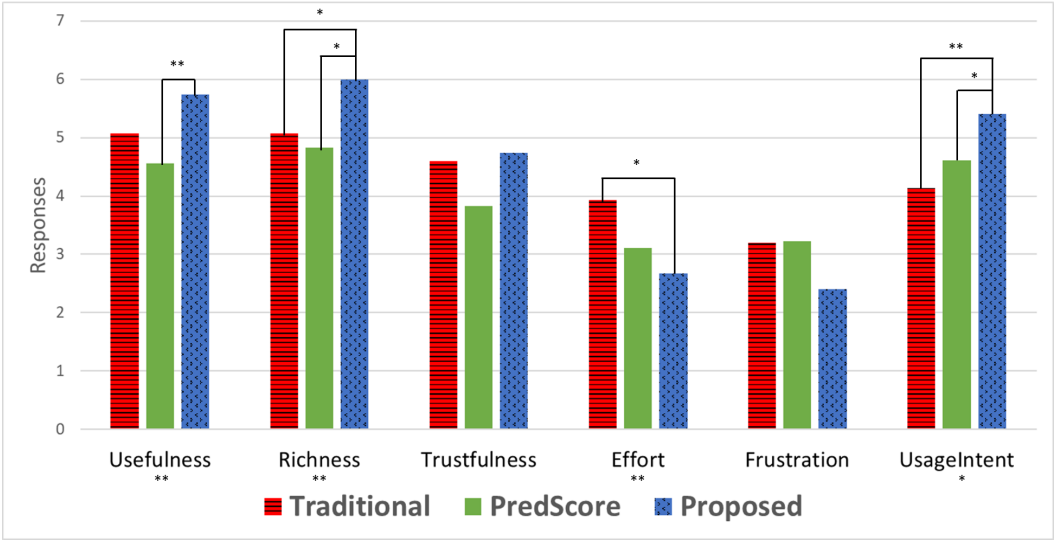


Fig. 5. Results of questionnaires on three interfaces. The proposed interface is more useful, richer, more trustful while reducing effort and frustration on assessment tasks. It is more likely to be used in the clinical practices than other baseline interfaces (i.e. traditional and predscore). We denote any 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% and 95% significance level.

The proposed interface achieves a **higher usefulness** score ($\mu = 5.73$) than the others (traditional: $\mu = 5.06$, $p = 0.15$ and predscore: $\mu = 4.55$, $p < 0.05$) as additional explanations of Condition 3 were considered *"useful to understand patient's condition"* for therapists. The proposed interface also receives a **significantly higher richness** score ($\mu = 6.00$) than the others (traditional: $\mu = 5.06$, $p < 0.10$ and proposed: $\mu = 4.83$, $p < 0.10$). As there is no statistical difference between traditional and predscore on a richness score (Table 7), this indicates the positive effect of user-specific analysis from the proposed interface on the richness. In addition, therapists expressed **higher trust** on the proposed interface ($\mu = 4.73$) than the others (traditional: $\mu = 4.60$ and

predscore: $\mu = 3.83$ without statistical significance). Although therapists identified that *"predicted scores of an interface are sometimes not matched with their own assessment and not trustful"*, analysis of the proposed interface complements to *"understand why such predicted scores are generated"*.

Therapists experienced **significantly lower effort** on assessment ($\mu = 2.66$) with the proposed interface than the others (traditional: $\mu = 3.93, p < 0.10$ and predscore: $\mu = 3.11, p = 0.27$). The predscore interface does not have a statistical difference with the traditional interface (Table 7). Thus, user-specific analysis of the proposed interface has a positive effect on lower effort. Also, the proposed interface has a **lower frustration** score ($\mu = 2.40$) than the others (traditional: $\mu = 3.20$ and predscore: $\mu = 3.22$ without statistical significance). Therapists described that user-specific analysis of the proposed interface (feature analysis, images of salient frames, and graphs of joint trajectories) reduce the effort and frustration to *"search evidence in videos"*. However, the frustration score of the proposed interface does not have any significant difference than that of other baseline interfaces.

In addition, the proposed interface 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 score between traditional and predscore is not statistically different (Table 7), the user-specific analysis of a system has a positive effect on the usage intent score. In term of preference responses, therapists mostly **prefer the usage of the proposed** interface to those of traditional and predscore interfaces:

Given the traditional and proposed interfaces,

- 2 out 7 therapists *'totally preferred'* the proposed interface to the traditional interface
- 4 out 7 therapists *'much more preferred'* the proposed interface to the traditional interface
- 1 out of 7 therapists *'much more preferred'* traditional interface to proposed interface

Given the predscore and proposed interfaces,

- 4 out of 7 therapists *'totally preferred'* the proposed interface to the predscore interface
- 2 out 7 therapists *'much more preferred'* the proposed interface to the predscore interface
- 1 out of 7 therapists *'slightly more preferred'* the predscore interface to the proposed interface

Overall, therapists commented that the proposed interface with predicted assessment and additional user-specific analysis *"is very interesting"* and *"gives me insights to assess patient's performance"*. The proposed interface achieved positive responses on all aspects of questionaires: our proposed interface provides more useful and richer information to understand the performance of a patient, leads to higher trust in the system, reduces therapist's efforts and frustration to find evidence for assessment, and is more likely to be used in clinical practice. However, only **usefulness**, **richness**, **effort**, and **usage intent** aspects have statistical significance using one-way ANOVA tests. When we compare score differences between the proposed and the traditional interfaces using post-hoc paired t-tests, **richness**, **effort**, and **usage intent** aspects have statistical significance.

8.2 Logs of an Interface

Figure 6a, 6b, and 6c describe the measurements of logs (i.e. time on analysis/video per assessment and an average number of video events) from three interfaces. We performed statistical analysis with the same procedure as described in the Section 8.1. We denoted any statistical significance using one-way ANOVA tests at the bottom of each variable. If a variable has any significance, we also indicated any statistically significant results using paired-tests at the top of each variable (Figure 6a, 6b, and 6c).

The proposed interface with additional user-specific analysis has **longer average time on assessment** ($\mu = 98.27$ seconds) than the traditional interface ($\mu = 79.43$) and the predscore interface ($\mu = 97.69$ seconds). When we analyze average time on videos, the proposed interface shows **lower average time on videos** ($\mu = 47.29$ seconds) than the others (the traditional interface:

$\mu = 79.43$ seconds and the predscore interface: $\mu = 81.97$ seconds). However, both average time on assessment and videos do not have any statistical significance using one-way ANOVA tests.

In addition, the proposed interface has a **significantly lower average number of video events, video playbacks** ($\mu = 4.25$) than the others (the traditional: $\mu = 6.45$ and the predscore: $\mu = 10.16$, $p < 0.01$). The proposed interface has the lowest average number of video playbacks. When we compare the traditional and predscore interfaces, the predscore interface has a higher average number of video playbacks. Thus, user-specific analysis of the proposed interface has an effect on lower number of video events.

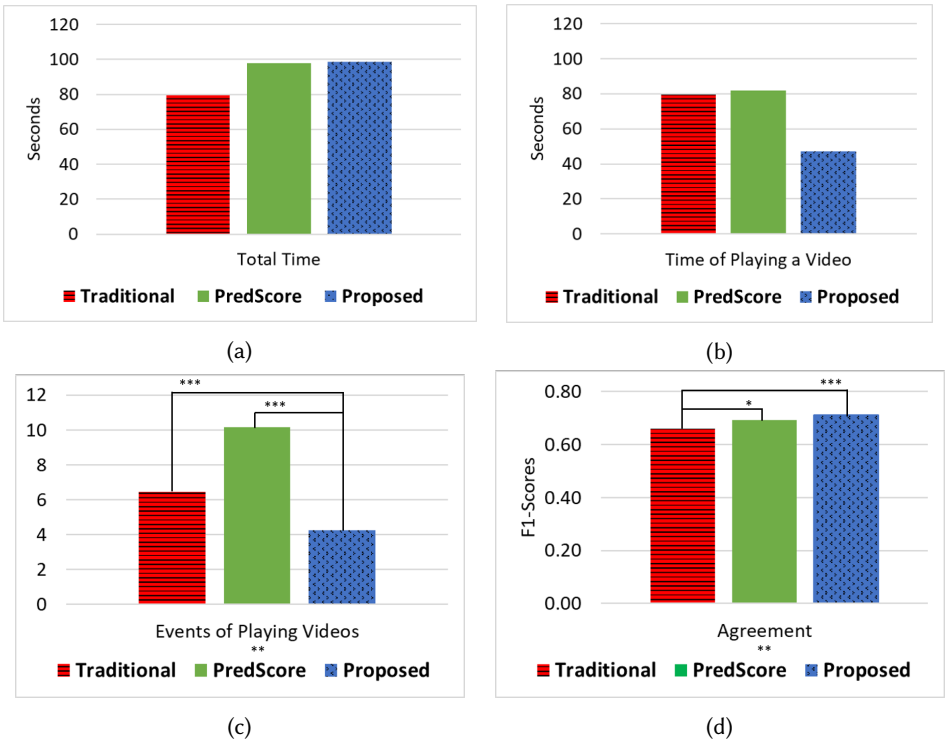


Fig. 6. Results of the user study in term of (a) total time on assessment, (b) total time on video, (c) an average number of video events, and (d) agreement level of therapists' assessment. Although the proposed interface has the longest total time of assessment, it has the shortest time on playing a video, the lowest number of video events, playbacks, and the highest agreement on therapists' assessment. We indicate any 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-test at 90% and 99% significance level.

8.3 Agreement Level of Therapists' Assessment

Figure 6d shows the agreement level of therapists' assessment using three interfaces. We performed statistical analysis with the same procedure as described in the Section 8.1, and indicated any statistical significance using one-way ANOVA test at the bottom of a variable and paired t-tests at the top of a variable in Figure 6d.

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

8.4 Post-study Interviews

After completing the user study, we collected therapists' perspectives on the proposed interface. Specifically, we asked their experiences and opinions on the interface and the possibility of accepting it in current practice.

8.4.1 Different Patterns to Utilize and Review Analysis.

Therapists described two different patterns of using the interface for the assessment. One pattern is to first *"review the feature analysis to get the overview of quantitative difference between unaffected and affected sides"* (TP 7) and *"validate quantitative feature analysis with images of salient frames and graphs of joint trajectories"* (TP 9). Some therapists preferred to get the initial insight of assessment from feature analysis, because *"quantitative, kinematic feature analysis between unaffected and affected sides is useful and fast to get insights and validate my assessment"* (TP 7). The other strategy is to first *"observe graphs of joint trajectories to understand the overview of a motion"* (TP 3) and *"review the detailed, quantitative feature analysis and images of salient frames"* (TP 8). Others reviewed graphs of joint trajectories first, because it *"provides various insights together (e.g. duration, amplitude, and tremor of a motion)"* (TP 3) to improve and validate therapist's initial, hypothetical assessment.

8.4.2 Understanding the Capabilities of a System to Adjust its Usage.

After reviewing predicted assessment and user-specific analysis, therapists were able to determine whether a system makes a mistake or not and understand the capabilities of a system. Even if the predicted scores of an interface sometimes mismatch with therapist's assessment, therapists consider *"the proposed interface is trustful"* (TP 9) in a way that *"I can review patient-specific analysis to understand whether a system fails to predict correctly or I make a mistake"* (TP 4). For example, TP 9 commented that *"the prediction on range of motion (ROM) seemed to be aligned most of the time with my hypothetical assessment and insights from user-specific analysis"*. In contrast, TP 9 mentioned that predictions of compensation do not sometimes perform well, because the system *"does not provide a prediction that is aligned with mine and analysis does not include leaning trunk to the side"* feature to predict compensation of a patient, who compensates trunk to the side.

Once therapists developed mental models on the capabilities of a system, they then adjusted their usage and validation on a system accordingly. They commented that they would trust more on system predictions of *'Range of Motion'* and less on system predictions on *'Compensation'* performance component. *"After finding more matching on predicted assessment on the range of motion (ROM), I spent less time to accept and validate predictions from a system on ROM and more time on validating predictions and analysis from a system to assess a compensation motion"* (TP 9). Thus, user-specific analysis of our interface assisted therapists to understand the capabilities of a system to predict performance components and develop the different levels of trust and usages on a system.

8.4.3 Benefits and Potentials of a System in Practice.
Overall, therapists considered the proposed system as *"a good platform"* (TP 5) for rehabilitation assessment. User-specific analysis of the interface (e.g. feature analysis, images of salient frames, and graphs of joint trajectories) *"brings more interesting, new aspects of a patient"* (TP 3) and enables therapists to *"understand why the predicted assessment is suggested"* (TP 9). Specifically, therapists found that feature analysis (Figure 3b) is *"easy and intuitive"* (TP 9) to *"quickly observe the quantitative difference between unaffected and affected sides"* (TP 6) for assessment. Images of salient frames (Figure 3c) *"was helpful to validate feature analysis"* (TP 9). Graphs of joint trajectories (Figure 3d) *"was useful to review the overall trends"* (TP 8) and *"the duration of a motion"* (TP 9), which were *"helpful to assess the smoothness of a motion"* (TP 4).

In addition, therapists experienced that user-specific analysis on each performance component is useful to decrease complexity of assessment and eventually *"reduces their efforts and frustration to search evidence on the assessment"* (TP 6). *"It is a complex and challenging process to simultaneously review multiple aspects of assessment while watching a video (TP 9)". "When only a video is presented"*, therapists have *"difficulty to consider different perspectives on assessment at the same time* (TP 9). So, therapists *"had to replay a video multiple times"* (TP 3). In contrast, *"user-specific analysis on each component from the interface simplified my assessment process"* (TP 9). *"I started reducing my effort to search clues for assessment through replaying a video, and relied more on analysis of the interface"*, because the interface *"quickly presents various quantitative measurements for assessment, which I have to speculate while watching a video"* (TP 9).

As the proposed interface is *"easy to use and quickly summarizes quantitative data with intuitive graphics to provide insights of patient's performance"* (TP 9), therapists are positive to accept the interface in practice. Therapists commented that currently they *"do not have much quantitative data to analyze and discuss with patients"* (TP 7). Predicted assessment and user-specific analysis with quantitative data from our interface could facilitate *"understanding on patient's performance and communicate it with patients"* (TP 4). In addition, some therapists considered that the interface might be *"helpful to motivate patient's participation in rehabilitation sessions"* (TP 9) by tracking and presenting the patient's progress with quantitative data.

9 DISCUSSION

In this work, we co-designed an intelligent decision support system with therapists and studied its effect on therapists' rehabilitation assessment. Our results show various effects of user-specific analysis, explanations of a system to supplement therapists' assessment. We discuss the importance of generating supplementary explanations on a system prediction with human-centered designs and making a system interactive for human and AI collaborative decision making in practice.

9.1 Effects of User-Specific Analysis, Explanations of a System

Instead of manually reviewing abundant features, our system with machine learning algorithms can automatically identify salient features of assessment to predict the quality of motion and generate succinct user-specific analysis. Our results show that this analysis can provide richer insights on the performance of a patient, but also serve as a source of information that allows therapists understanding the capabilities and limitations of a system [9]. Specifically, after reviewing user-specific analysis, therapists identified that our system is more competent in assessing the *'Range of Motion'* performance component than the *'Compensation'* performance component. Furthermore, therapists could enumerate a potential reason for a system failure (e.g. not including an important feature of assessment) and asked the possibility of updating a system. Overall, therapists did not blindly utilize the information of a system. Instead, they developed a mental model on the competency of a system through analyzing user-specific analysis and exploited a system accordingly.

While reviewing user-specific analysis, therapists spent longer time on assessment using the proposed interface than the traditional interface. However, longer time on overall assessment does not necessarily mean the degraded user-experience. In spite of longer assessment time on the proposed interface, therapists provided significantly lower scores on their efforts to assess patient's exercises with the proposed interface than the traditional interface (Figure 5). Therapists considered reviewing explanations, user-specific analysis is helpful to *"understand patient's performance and validate my/therapist's assessment"* (TP 7) than repetitively watch a video for assessment. In addition, they provided significantly higher usage intent scores on the proposed interface than the traditional interface (Figure 5). The proposed interface with user-specific analysis supported significantly more consistent assessment of therapists (Figure 6d). Thus, a system with salient explanations is more likely to be adopted in practice while improving expert's performance on a task.

9.2 Considerations for a Decision Support System

Although the full interpretation of a machine learning model is still challenging, our study showed that feature selection can be utilized to identify salient aspects of a decision making task for deriving a machine learning model of a task and providing a means of communication with experts (i.e. user-specific analysis of our system). When a means of communication is developed with feature selection, we found that an early involvement of experts in a design phase is necessary and beneficial to provide contextually relevant information with understandable terms.

Another consideration is to make an adaptive system for personalization and collaborative decision making between an AI-based system and experts. During the study, we observed that therapists developed different usage patterns of a decision support system and utilized different functionalities or information of a system based on a task and their own knowledge. Some therapists preferred to get the initial insight of assessment from feature analysis and others reviewed trajectory analysis first (Section 8.4). Thus, making a system adaptable to each therapist's preference is recommended so that therapists can quickly collect necessary information, evidence for their decision making in practice.

In addition, TP 6 suggested that it would be useful if a therapist can tune a system by including or excluding an identified feature to utilize a different set of features based on an individual therapist's experience and correct any mismatched prediction scores of a system. For instance, when the interface failed to predict correctly the assessment on compensation (e.g. when a patient leaned trunk to the side), TP 9 noticed that *"leaning trunk to the side"* feature was not included in user-specific analysis and asked whether a system can be updated to include this feature for the prediction on *'Compensation'* performance component. We believe that if a system repetitively makes a mistake and does not get better, experts might end up having lower acceptance on such a system. Thus, designers should consider applying an interactive technique [12, 32, 37] to make a system adaptive for better personalization and integration into clinical practice. A promising direction of future research is to explore how human and machine intelligence can complement each other to improve a complex decision making task [37].

9.3 Potential Impact and Limitations

Even if various medical disciplines rely on standardized guidelines [17, 25, 56, 59], uncertainty is inescapable in practice [20]. As experts often rely on their experience-based heuristics, disagreement on decision making is prevalent [4, 30, 31]. Our results show that a decision support system with a machine learning algorithm can replicate a therapist's assessment with decent performance that is comparable to therapist's agreement (Table 4) . In addition, this system can provide alternative perspectives with quantitative analysis on a decision making task and support experts more

consistent decision making (Figure 6d). We believe such a decision making system has the potential to be an assistant of experts, but also a learning tool to reduce distorting decision making [20].

Although our results demonstrate the potential of our decision support system, one of the limitations of this study is the small sample size of therapists in evaluation: seven therapists from four rehabilitation centers do not represent all therapists. However, such small sample size is not unusual among similar studies [8]. In addition, as this study mainly evaluated only one decision making task (i.e. rehabilitation assessment), expansion to other decision making tasks and additional validation during an extended period are necessary for further generalization.

10 CONCLUSION

In this paper, we co-designed, developed, and evaluated a machine learning-based decision support system for stroke rehabilitation assessment. This system automatically identifies salient features of assessment to predict the quality of motion and generate summarized user-specific analysis as explanations on its prediction. According to the evaluation with therapists, we found that the presentation of predicted assessment with salient explanations leads to higher usage intention of therapists while bringing them richer insights on the performance of a patient, reducing their effort on assessment, and supporting them to have more consistent assessment. This work highlights the importance of generating salient explanations on predictions of a clinical decision support system with human-centered designs and making a system interactive for better deployment in practice.

A APPENDIX

A.1 Implementation Details of Reinforcment Learning

We implement a neural network model (Q_θ) for deep Q-learning [37, 60] using *'PyTorch'* libraries [47]. The input layer of the network consists of feature and binary mask vectors. This masking input vector is to indicate whether a feature is recruited or not [24]. Specifically, we 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.

For training a model, we take a batch of transitions that are empirically experienced by the agent with a greedy policy $\pi_\theta(s) = \max_a Q_\theta(s, a)$, 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 received reward and γ indicates the discounted factor. We clip a gradient if a gradient norm exceeds 1.0 [24] and update the target network after each step. Instead of directly updating the weight of the target network, we apply soft target updates [37]: $\theta' \leftarrow \rho\theta + (1 - \rho)\theta'$, 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 target networks. As the application of soft target updates may lead to slow learning, we apply an experience replay [37] 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 ϵ_{start} (0.5) to the ϵ_{end} (0.05) with a step value, ϵ_{step} (0.02).

A.2 Other Details of Experiment, Implementation, and Analysis

Table 5. Profiles of 15 Post-Stroke Patients

Patient ID	Total Fugl (0-66)	Age	Sex	Affected Side	Type
P01	65	69	M	Left	Not Specified
P02	65	60	M	Left	Hemorphagic
P03	66	61	M	Left	Not Specified
P04	66	63	M	Right	Ischemic
P05	55	51	M	Left	Ischemic
P06	13	63	M	Left	Ischemic & Spastic
P07	42	86	F	Right	Ischemic
P08	15	71	M	Left	Ischemic
P09	35	78	M	Left	Hemorrphagic
P10	21	53	M	Right	Ischemic
P11	16	37	M	Right	Ischemic
P12	11	61	M	Left	Hemorrphagic
P13	46	59	M	Left	Ischemic
P14	11	67	M	Left	Ischemic
P15	34	66	F	Left	Ischemic

Table 6. 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 7. Statistical Analysis (Anova and Pairwise T-Test) on the Results of the User Study

	Responses						Time		Events	Agreement
	Usefulness	Richness	Trust	Effort	Frust	UsageIntent	Overall	Video		
Anova Tests	**p < 0.05**	**p < 0.05**	p = 0.19382	**p < 0.05**	p = 0.1962	**p < 0.10**	p = 0.3183	p = 0.2582	**p < 0.05**	**p < 0.05**
Traditional vs PredScore	p = 0.1275	p = 0.3576	n/a	p = 0.1290	n/a	p = 0.2154	n/a	n/a	**p < 0.01**	**p < 0.10**
Traditional vs Proposed	p = 0.1514	**p < 0.10**	n/a	**p < 0.10**	n/a	**p < 0.05**	n/a	n/a	**p < 0.01**	**p < 0.01**
PredScore vs Proposed	**p < 0.05**	**p < 0.10**	n/a	p = 0.2714	n/a	**p < 0.10**	n/a	n/a	**p < 0.01**	p = 0.1874

ACKNOWLEDGMENTS

The authors thank all the participants in this study for their dedication, time and valuable inputs. Also, we thank Aaron Steinfeld for his valuable comments on experimental designs and the anonymous reviewers for their constructive comments and suggestions on the manuscript. This work is partially supported by the IntelligentCare project (LISBOA-01-0247-FEDER-045948), co-financed by the European Regional Development Fund (ERDF) through the LISBOA 2020 and the FCT under CMU-PT. Additional support was provided by the FCT [SFRH/BD/113694/2015, LARSyS - Plurianual funding 2020-2023 (UIDB/50009/2020)], CMU GuSH Research Grant, and the National Science Foundation (NSF) under grant number CNS-1518865.

REFERENCES

[1] 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.

[2] PK Anooj. 2012. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences* 24, 1 (2012), 27–40.

[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] 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).

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] Chien-Yen Chang, Belinda Lange, Mi Zhang, Sebastian Koenig, Phil Requejo, Noom Somboon, Alexander A Sawchuk, and Albert A Rizzo. 2012. Towards pervasive physical rehabilitation using Microsoft Kinect. In *2012 6th international conference on pervasive computing technologies for healthcare (PervasiveHealth) and workshops*. IEEE, 159–162.

[11] Maria do Carmo Vilas-Boas and João Paulo Silva Cunha. 2016. Movement quantification in neurological diseases: methods and applications. *IEEE reviews in biomedical engineering* 9 (2016), 15–31.

[12] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 39–45.

[13] Juan Fasola and Maja J Matarić. 2013. A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction* 2, 2 (2013), 3–32.

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

[15] Nicola K Gale, Gemma Heath, Elaine Cameron, Sabina Rashid, and Sabi Redwood. 2013. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC medical research methodology* 13, 1 (2013), 117.

[16] Amit X Garg, Neill KJ Adhikari, Heather McDonald, M Patricia Rosas-Arellano, Philip J Devereaux, Joseph Beyene, Justina Sam, and R Brian Haynes. 2005. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama* 293, 10 (2005), 1223–1238.

[17] 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.

[18] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA)* 2 (2017).
[19] 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.
[20] Katherine H Hall. 2002. Reviewing intuitive decision-making and uncertainty: the implications for medical education. *Medical education* 36, 3 (2002), 216–224.
[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] Henk T Hendricks, Jacques van Limbeek, Alexander C Geurts, and Machiel J Zwarts. 2002. Motor recovery after stroke: a systematic review of the literature. *Archives of physical medicine and rehabilitation* 83, 11 (2002), 1629–1637.
[23] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.
[24] Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. 2019. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3959–3966.
[25] Joseph Jankovic. 2008. Parkinson's disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry* 79, 4 (2008), 368–376.
[26] Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. 2005. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj* 330, 7494 (2005), 765.
[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] M. H. Lee. 2019. Intelligent Agent for Assessing and Guiding Rehabilitation Exercises. In *IJCAI*. 6444–6445.
[34] Min Hun Lee. 2019. An Intelligent Decision Support System for Stroke Rehabilitation Assessment. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 694–696.
[35] 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.
[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] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Towards Personalized Interaction and Corrective Feedback of a Socially Assistive Robot for Post-Stroke Rehabilitation Therapy. In *Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication*.
[39] Ho Shing Lo and Sheng Quan Xie. 2012. Exoskeleton robots for upper-limb rehabilitation: State of the art and future prospects. *Medical engineering & physics* 34, 3 (2012), 261–268.
[40] 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.
[41] Rui CV Loureiro, William S Harwin, Kiyoshi Nagai, and Michelle Johnson. 2011. Advances in upper limb stroke rehabilitation: a technology push. *Medical & biological engineering & computing* 49, 10 (2011), 1103.
[42] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012).
[43] Karina Iglesia Molina, Natalia Aquaroni Ricci, Suzana Albuquerque de Moraes, and Monica Rodrigues Perracini. 2014. Virtual reality using games for improving physical functioning in older adults: a systematic review. *Journal of*

neuroengineering and rehabilitation 11, 1 (2014), 156.
[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] Marijke Rensink, Marieke Schuurmans, Eline Lindeman, and Thora Hafsteinsdottir. 2009. Task-oriented training in rehabilitation after stroke: systematic review. *Journal of advanced nursing* 65, 4 (2009), 737–754.
[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).
[52] M Joan Saary. 2008. Radar plots: a useful way for presenting multivariate health care data. *Journal of clinical epidemiology* 61, 4 (2008), 311–317.
[53] Dylan M Smith, Stephanie L Brown, and Peter A Ubel. 2008. Mispredictions and misrecollections: challenges for subjective outcome measurement. *Disability and Rehabilitation* 30, 6 (2008), 418–424.
[54] Robert L Spitzer, Janet BW Williams, and Jean Endicott. 2012. Standards for DSM-5 reliability. *American Journal of Psychiatry* 169, 5 (2012), 537–537.
[55] Chuan-Jun Su, Chang-Yu Chiang, and Jing-Yan Huang. 2014. Kinect-enabled home-based rehabilitation system using Dynamic Time Warping and fuzzy logic. *Applied Soft Computing* 22 (2014), 652–666.
[56] 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.
[57] David H Sutherland. 2002. The evolution of clinical gait analysis: Part II Kinematics. *Gait & posture* 16, 2 (2002), 159–179.
[58] Jiliang Tang, Salem Alelyani, and Huan Liu. 2014. Feature selection for classification: A review. *Data classification: Algorithms and applications* (2014), 37.
[59] Edward Taub, David M Morris, Jean Crago, Danna Kay King, Mary Bowman, Camille Bryson, Staci Bishop, Sonya Pearson, and Sharon E Shaw. 2011. Wolf motor function test (WMFT) manual. *Birmingham: University of Alabama, CI Therapy Research Group* (2011).
[60] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*.
[61] Jasper E Visser, Mark G Carpenter, Herman van der Kooij, and Bastiaan R Bloem. 2008. The clinical utility of posturography. *Clinical Neurophysiology* 119, 11 (2008), 2424–2436.
[62] 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).
[63] 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.
[64] 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.
[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*.
[66] Huiyu Zhou and Huosheng Hu. 2008. Human motion tracking for rehabilitation—A survey. *Biomedical Signal Processing and Control* 3, 1 (2008), 1–18.

Received January 2020; revised June 2020; accepted July 2020