User Modeling and User-Adapted Interaction
https://doi.org/10.1007/s11257-022-09348-5

Design, development, and evaluation of an interactive personalized social robot to monitor and coach post-stroke rehabilitation exercises

Min Hun Lee[1] · Daniel P. Siewiorek[2] · Asim Smailagic[2] · Alexandre Bernardino[3] · Sergi Bermúdez i Badia[4]

Received: 17 January 2021 / Accepted in revised form: 22 October 2022

Abstract

Socially assistive robots are increasingly being explored to improve the engagement of older adults and people with disability in health and well-being-related exercises. However, even if people have various physical conditions, most prior work on social robot exercise coaching systems has utilized generic, predefined feedback. The deployment of these systems still remains a challenge. In this paper, we present our work of iteratively engaging therapists and post-stroke survivors to design, develop, and evaluate a social robot exercise coaching system for personalized rehabilitation. Through interviews with therapists, we designed how this system interacts with the user and then developed an interactive social robot exercise coaching system. This system integrates a neural network model with a rule-based model to automatically monitor and assess patients' rehabilitation exercises and can be tuned with individual patient's data to generate real-time, personalized corrective feedback for improvement. With the dataset of rehabilitation exercises from 15 post-stroke survivors, we demonstrated our sys-

✉ Min Hun Lee
mhlee@smu.edu.sg

Daniel P. Siewiorek
dps@cs.cmu.edu

Asim Smailagic
asim@cs.cmu.edu

Alexandre Bernardino
alex@isr.tecnico.ulisboa.pt

Sergi Bermúdez i Badia
sergi.bermudez@uma.pt

[1] Singapore Management University, Singapore, Singapore

[2] Carnegie Mellon University, Pittsburgh, USA

[3] Instituto Superior Técnico, Lisbon, Portugal

[4] NOVA-LINCS, University of Madeira, Funchal, Portugal

Published online: 11 March 2023

tem significantly improves its performance to assess patients' exercises while tuning with held-out patient's data. In addition, our real-world evaluation study showed that our system can adapt to new participants and achieved 0.81 average performance to assess their exercises, which is comparable to the experts' agreement level. We further discuss the potential benefits and limitations of our system in practice.

Keywords Human–robot interaction · Socially assistive robots · Personalization · Post-stroke rehabilitation therapy

1 Introduction

As the world's older population continues to grow at an unprecedented rate, the current supply of care providers is insufficient to meet the current and ongoing demand for care services (Dall et al. 2013). Researchers have explored an opportunity of socially assistive robots (Feil-Seifer et al. 2005; Tapus and Mataric 2006) that aim to enable people with cognitive, sensory, and motor impairments or assist the clinical workforce (Riek 2017). One potential application is socially assistive robots for rehabilitation therapy (Matarić et al. 2007; Lee et al. 2020, 2022). During rehabilitation, patients require completing a significant amount of self-directed exercises (O'Sullivan et al. 2019; Lee et al. 2022). However, low treatment adherence is a problem across several healthcare disciplines of physiotherapy (Kåringen et al. 2011). To address these problems, there has been increasing attention on social robot coaching systems (Riek 2017; Matarić et al. 2007; Lee et al. 2020, 2022). These systems autonomously monitor patients' exercises and provide encouragement to support patients' engagement in well-being-related or rehabilitation exercises through social interaction (Tapus et al. 2007; Feil-Seifer et al. 2005).

Prior work on robotic exercise coaching systems has demonstrated that older adults or post-stroke subjects can successfully exercise and stay engaged with a robot over sessions (Fasola and Matarić 2013; Görer 2013). However, despite the potential of a robot to monitor and guide exercises, prior work is limited to providing generic, predefined corrective feedback on patient's exercise performance (e.g., checking angular difference with the prespecified motion (Görer 2013; Fasola and Matarić 2013; Guneysu and Arnrich 2017)). It is still challenging to empower a social robot exercise coaching system to generate tailored corrective feedback on an individual patient's motion (Görer 2013) and adopt these systems broadly yet.

In this work, we design, develop, and evaluate a socially assistive robot coaching system that automatically monitors and coaches physical rehabilitation therapy. Specifically, we selected a test domain as stroke, which is the second leading cause of death and disability (Feigin et al. 2017). We then conducted interviews with therapists to design and develop a socially assistive robot coaching system. This system integrates a machine learning (ML) model with a rule-based (RB) model and can be tuned with held-out user data to assess the performance of exercises for personalized post-stroke therapy (Fig. 1a) (Lee et al. 2020). Building upon the previous work (Lee et al. 2020), we demonstrated the benefit of our approach to adapt a new user and provide personalized assessment compared to the commonly used transfer learning

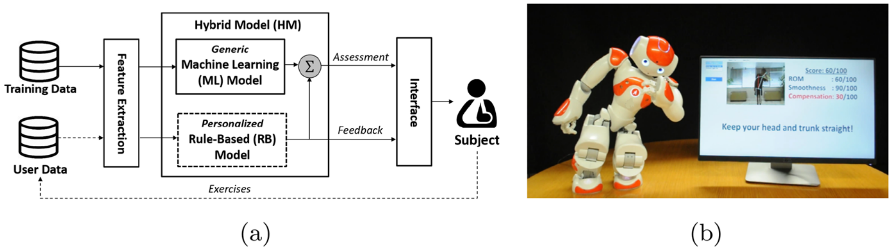

Fig. 1 **a** Flow diagram of an interactive approach of a socially assistive robot for personalized physical therapy. **b** a setup of the system with a visualization interface and a socially assistive robot that provides corrective feedback (e.g., audio, visual, gestures of the robot)

technique on a feed-forward neural network model (Zhuang et al. 2020) (i.e., pretrains a model using the dataset from post-stroke survivors and then fine-tune it based on the data from a new post-stroke survivor).

During the real-world study with ten participants, our interactive system can be adapted to new participants and achieved 0.81 average performance to assess participants' quality of motion, which is comparable to experts' agreement level (i.e., 0.80 average performance). Overall, participants expressed positive opinions on our system to monitor and provide feedback on their exercises, but also described practical issues to be improved.

2 Related work

In this section, we describe the background of socially assistive robotics for coaching exercises and outline related work on designs and techniques of socially assistive robotics for rehabilitation therapy.

2.1 A socially assistive robot as a coach

The research of socially assistive robotics has shown great potential to supplement healthcare services through social interaction (Feil-Seifer et al. 2005; Tapus et al. 2009; Matarić and Scassellati 2016). For instance, researchers have explored the feasibility of a socially assistive robot exercise coaching system in a rehabilitation process, in which the system automatically monitors rehabilitation exercises and provides users feedback without the presence of a therapist (Matarić et al. 2007). Fasola and Mataric demonstrated that older adults considered a physically embodied robot more engaging and acceptable as an exercise partner than a virtually embodied agent (Fasola and Matarić 2013). Furthermore, researchers have shown that diverse populations (i.e., post-stroke patients (Fasola and Matarić 2013), elderly people (Görer et al. 2017), children (Guneysu and Arnrich 2017)) can engage in exercise sessions with a social robot exercise coaching system on several domains (e.g., stroke, dementia, etc.) (Tapus et al. 2009; Lee et al. 2022; Riek 2017).

2.2 Designs of a socially assistive robot coaching system

For creating user-centered, socially assistive robots, researchers have engaged with various stakeholders to derive design requirements (Beer et al. 2012; Azenkot et al. 2016; Winkle et al. 2018; Lee et al. 2022). Winkle et al. (2018) described design guidelines of social robots for rehabilitation, from focus group sessions and interviews with therapists. In addition, Polak and Levy-Tzedek also conducted focus group sessions with therapists and a preliminary evaluation study on a gamification system for rehabilitation with four post-stroke survivors (Feingold Polak 2020). Lee et al. (2022) conducted studies with therapists and post-stroke survivors to elicit detailed design specifications on how AI and robotic coaches could interact with and guide patients' exercises in an effective and acceptable way. One of the important design considerations that are repetitively mentioned in prior work is the importance of personalized feedback (Winkle et al. 2018; Feingold Polak 2020; Lee et al. 2022).

In this work, we interviewed therapists to understand what kinds of feedback they generate and explored a computational technique that enables a social robot exercise coaching system to generate personalized feedback and control robot behaviors as a therapist.

2.3 Techniques of monitoring and assessing patient's exercises

The capability of automatically assessing a patient's motion and providing a personalized interaction with tailored corrective feedback on patient's exercise performance is critical for the deployment of a social robot exercise coaching system (Matarić et al. 2009; Görer et al. 2017; Lee et al. 2022). For personalized interactions with a socially assistive robot, Irfan et al. explored recognizing a user and referred the user's name periodically as personalized feedback (Irfan et al. 2020). Schneider and Kummert investigated a technique to match the user's preferred order of different exercises for personalized interactions of exercise robots (Schneider and Kummert 2021). However, limited prior work on social robot exercise coaching systems has explored how an automated assessment approach can be developed to generate personalized corrective feedback.

When it comes to an automated assessment approach, researchers have implemented a method that monitors the completion of an exercise by computing the difference of a joint angle between the user's motion and the predefined target motion (Fasola and Matarić 2013; Görer et al. 2017). Guneysu and Arnrich (2017) applied dynamic time warping to compute the statistics of a joint angle and distance measures with a predefined motion. Tanguy et al. (2016) utilized a Gaussian mixture model to generate an ideal motion and arbitrarily set a threshold value to identify the differences of joints between idea and observed motions. Although both (Guneysu and Arnrich 2017) and (Tanguy et al. 2016) support analyzing multiple variables for evaluating an exercise, they still rely on a predefined motion or a generic threshold. Prior work with generic threshold-based methods might not be applicable for patients with various characteristics (Lee et al. 2020).

In addition, researchers have also explored a machine learning model to monitor patients' quality of motion. For instance, Kashi et al. (2020) evaluated the feasibility of a random-forest model to identify compensatory movements. However, it remains unclear how such a machine learning model can adapt to a new patient and perform well to assess patients' quality of motion.

For personalized quantitative rehabilitation assessment, Lee et al. (2020) explored an approach of dynamic feature selections and a hybrid model that integrates a machine learning model with a rule-based model (Lee et al. 2020). However, prior work is limited to providing assessment after completing a motion and does not support frame-level assessment to provide any information on when an erroneous motion has occurred.

Building upon prior work that explores a hybrid model for personalized assessment (Lee et al. 2020, ?), we further investigated the system implementation of a socially assistive robot to automatically monitor and guide patients' exercises. Specifically, we analyzed the benefit of our interactive hybrid approach compared to the commonly used transfer learning technique on a feedforward neural network model (Zhuang et al. 2020; Weiss et al. 2016) (i.e., pretrains the model using the dataset from other post-stroke survivors and then fine-tune it based on data from a new post-stroke survivor). In addition, we conducted a real-world experiment to evaluate the feasibility to adapt to a new participant and provide personalized, real-time corrective feedback.

3 Study for stroke rehabilitation

This work focuses on the domain of stroke, which is the second leading cause of death and third most common contributor to disability (Feigin et al. 2017). First, we iteratively discussed with three therapists (TPs with check marks in the specification column of Table 1; mean (M) = 6.33, standard deviation (SD) = 2.05 years of experience in stroke rehabilitation) to specify the study designs on stroke rehabilitation: exercises and performance components for assessment (Lee et al. 2019). We then had additional interviews with therapists to learn their practices on how they guide rehabilitation assessment. Based on these interviews, we created an interactive social robot coaching system that automatically monitors and coaches rehabilitation exercises. We then conducted a real-world experiment with ten healthy participants to evaluate the potential benefits and limitations of our system. This section presents only the specifications of our study and interviews with therapists to understand their practices. The evaluation part will be discussed later in Sect. 6.2 after presenting our system implementation.

3.1 Three task-oriented upper-limb exercises

This work utilizes three upper-limb stroke rehabilitation exercises recommended by therapists (Lee et al. 2020). For Exercise 1, a user has to raise the user's wrist to the mouth as if drinking water (Fig. 2a). For Exercise 2, a user has to raise the user's wrist

Table 1 Profiles of therapists, who contributed to specify the study and share their practices to design our system

ID	Specification	Interview	# of Years in stroke rehabilitation
TP 1	✓	✓	6
TP 2	✓	✓	4
TP 3	✓		9
TP 4		✓	23

forward as if touching a light switch on the wall (Fig. 2c). For Exercise 3, a user has to extend the user's elbow in the seated position to practice the usage of a cane (Fig. 2e).

3.2 Unaffected and affected sides

When a stroke occurs, post-stroke survivors suffer from the paralyzed, limited functional abilities of limbs. In this work, we refer to the unparalyzed side of a post-stroke survivor as the *"Unaffected"* side and the paralyzed side of a post-stroke survivor with limited functional ability as the *"Affected"* side.

3.3 Performance components

We discussed commonly used stroke assessment tools (i.e., the Wolf Motor Function Test (Wolf et al. 2001) and the Fugl–Meyer Assessment (Sanford et al. 1993)) with therapists and specified three common performance components of stroke rehabilitation exercises: *'Range of Motion (ROM)'*, *'Smoothness'*, and *'Compensation'* (Lee et al. 2020). The *'ROM'* indicates how closely a patient performs the target position of a task-oriented exercise. The *'Smoothness'* describes the degree of trembling and irregular movement of joints while performing an exercise. The *'Compensation'* indicates whether a patient performs any unnecessary, compensatory movements to achieve a target movement. For instance, patients might lean their head or trunk to the side and elevate their shoulder to perform an exercise using their affected side with the limited functional ability (Fig. 2).

The descriptions and labels of performance components are described in Table 2. The labels of *'ROM'* and *'Smoothness'* are annotated at the end of a motion and represented as a binary label on each performance component: a correct/normal performance component ($Y = 1$) and an incorrect/abnormal performance component ($Y = 0$). The labels of *'Compensation'* are annotated at every frame of the patient's motion to indicate whether three major compensations (i.e., abnormal alignment of head, spine, and shoulder) occur or not.

3.4 Understanding therapists' practices

We conducted a one-on-one interview with each of the three therapists (TPs with check marks in the interview column of Table 1; mean (M) = 11.00, standard deviation (SD) = 8.52 years of experience in stroke rehabilitation). Dur-

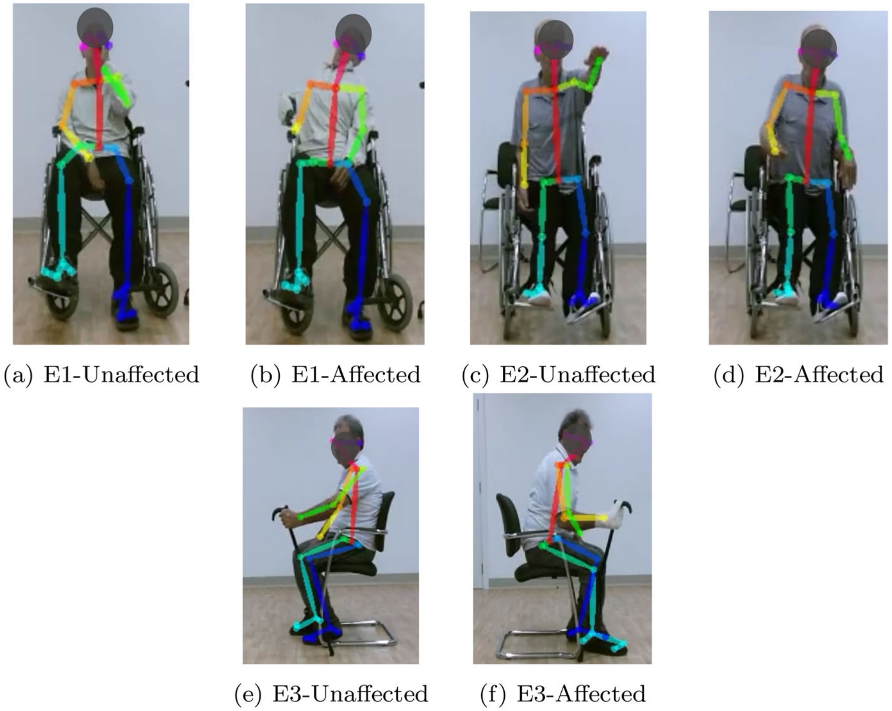

Fig. 2 Sample unaffected and affected motions of exercises: **a** a patient raises the patient's unaffected side of the wrist to the mouth, **b** a patient compensated with trunk and shoulder joints when attempting to move the patient's affected side of the wrist, **c** a patient raises patient's unaffected side of the wrist forward, **d** a patient elevated shoulder to compensate the limited functional ability of the patient's affected side, **e** a patient extends the patient's affected side of the elbow, and **f** a patient leaned trunk forward to extend the patient's elbow

Table 2 Performance components and labels of physical stroke rehabilitation exercises

Performance components	Labels	Guidelines
Range of motion (ROM)	0	Movement that does not achieve a *'Target'* position
	1	Movement achieves a *'Target'* position
Smoothness	0	Movement with tremor or unsmooth coordination
	1	Smoothly coordinated movement
Compensation	0/1	Head in abnormal/normal alignment
	0/1	Spine in abnormal/normal alignment
	0/1	Shoulder in abnormal/normal alignment

ing the 1-h interview, the researcher asked therapists to speak aloud their strategies for coaching a rehabilitation session and providing feedback on patient's exercises (i.e., *"what kinds of feedback do you generate for a post-stroke survivor?"*). To assist therapists' speaking aloud process, the researcher showed them the videos of post-stroke survivors, who have different functional abilities (i.e., high, moderate, and low capability to achieve an exercise) and perform rehabilitation exercises. The detailed process of collecting these videos is described in Sect. 5.1.

We analyzed the transcripts of interviews with therapists through an iterative coding process (Gale et al. 2013). Specifically, we first open-coded interview transcripts, discussed emerging themes and ideas, and iteratively improved our codebook. We found that therapists oversee the treatments of a post-stroke survivor by providing a personalized rehabilitation session. Specifically, they monitor how their patients perform an exercise and provide their patients feedback to support the correct execution of an exercise and encourage participation in rehabilitation (Lee et al. 2022). For guiding a rehabilitation session, we noticed that all three therapists have a simple and common procedure (Lee et al. 2022). Specifically, when they start a session, they engage with their patients through brief greetings and describe the goal of a session (e.g., what kinds of exercises a patient will perform and how many repetitions are recommended) (Lee et al. 2022). If a patient is not familiar with an exercise motion, therapists might show themselves to instruct a motion that a patient has to practice (Lee et al. 2022). When a patient performs an exercise, therapists monitor the patient's exercises to identify any part for improvement and provide corrective feedback (Lee et al. 2022). For instance, we found that therapists are particularly attentive to providing feedback on compensatory motions (Cirstea and Levin 2000) that might cause more severe pains. As rehabilitation requires patient engagement over an extended period, therapists also strive to provide positive encouragement to their patients (Lee et al. 2022).

4 Interactive approach of an socially assistive robot for personalized assessment and feedback

This work presents an interactive approach of a social robot exercise coaching system (Fig. 1a), which combines machine learning (ML) and rule-based (RB) models to assess the performance of patient's exercises and tunes with patient's data to generate personalized feedback (Lee et al. 2020). An ML model of our approach aims to extract meaningful patterns from a large amount of data and to support a generic assessment of the patient's quality of motion (Lee et al. 2020). As such, a generic ML model might not perform well on an unobserved new patient's motion with unique characteristics; our approach also integrates an ML model with a personalized RB model that can tune with the patient's unaffected motions to derive patient-specific threshold values. This RB model can be easily updated to complement a generic ML model through a weighted average, ensemble technique (Lee et al. 2020) into a hybrid model (HM) and utilized to generate personalized corrective feedback on patient's exercises. To provide feedback when an erroneous motion has occurred, we explored an ensemble voting method that leverages predictions on multiple consecutive frames for a more accurate frame-level assessment (Lee et al. 2020). In the following subsections, we describe the components of our approach: feature extraction, ML models, RB models, hybrid models, an ensemble voting method, and an interface of a socially assistive robot for personalized rehabilitation therapy.

4.1 Feature extraction

This work represented an exercise motion with sequential joint coordinates from a Kinect v2 sensor (Microsoft, Redmond, USA) and extracted various kinematic features (Lee et al. 2019). For the *'ROM'* performance component, we computed joint angles (e.g., elbow flexion, shoulder flexion, elbow extension), the distance to a target position, and normalized relative joint trajectories (i.e., the Euclidean distance between two joints—head and wrist, head and elbow) (Lee et al. 2019). For the *'Smoothness'* performance component, we computed the following speed-related features: speed and the zero-crossing ratio of acceleration (Lee et al. 2019). As our work focuses on upper-limb exercises, we computed these speed-related features on wrist and elbow joints. For the *'Compensation'* performance component, we computed normalized joint trajectories: distances between joint positions of the head, spine, and shoulder in x, y, z axis from the initial to the current frame (Lee et al. 2019).

A moving average filter with a window size of five frames was applied to reduce the noise of the estimated joint positions from a Kinect sensor similar to Lee et al. (2019). Given an exercise motion, we computed a feature matrix $\mathbf{F} = \{f_1, \ldots, f_T\} \in R^{T \times d}$ with T number of frames and d features and the statistics (e.g., maximum, minimum, range, average, and standard deviation) of a feature matrix over all frames of the exercise to summarize a motion into a feature vector, $X \in R^{5d}$. This summarized feature vector was utilized for the assessment of *'ROM'* and *'Smoothness'* performance components. In addition, unlike (Lee et al. 2019) that only supports offline assessment on the *'Compensation'* performance component, we extracted a feature vector at each frame for real-time, frame-level assessment on the *'Compensation'* performance component. Overall, we extracted 30 features for the *'ROM'* performance component, 60 features for the *'Smoothness'* performance component, and 9 features for the *'Compensation'* performance component.

4.2 Machine learning (ML) model

For a machine learning (ML) model, we applied a supervised learning algorithm through leave-one-patient-out cross-validation that utilizes training data from all patients except a patient for testing. The ML model computes the score of being correct on a performance component (P_{ML}) and predicts the quality of motion. Among various supervised learning algorithms, a decision tree, linear regression, a support vector machine, a feedforward neural network, and a long short-term memory (LSTM) network, we utilized a feedforward neural network (NN) model due to its outperformance as shown in Lee et al. (2020). For the implementation of a feedforward neural network (NN) model, we explored various architectures (i.e., one to three layers with 32, 64, 128, 256, and 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 applied 'ReLu' activation functions and 'AdamOptimizer' and trained a model with cross-entropy loss and a mini-batch size of 1 and an epoch of 1.

4.3 Rule-based (RB) model

A rule-based (RB) model leverages the set of feature-based, *if-then* rules from therapists to estimate the quality of a motion (Lee et al. 2020). In addition, our system computes statistics of kinematic features from user data and generates patient-specific rules for personalized assessment. For the initial development of the RB model, semi-structured interviews were conducted with two therapists (mean (M) = 5.05, standard deviation (SD) = 1.05 years of experience in stroke rehabilitation) to elicit their knowledge of assessing stroke rehabilitation exercises. The knowledge of therapists has been formalized as 15 independent *if-then* rules (Appendix Table 5). For example, the assessment on the ROM component for Exercise 1 is specified as follows (Lee et al. 2020):

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

where $p(j, c)$ indicates the joint position (p) with a joint j (e.g., wrist (wr) and spine shoulder, the top of spine, (spsh)) and the coordinate of a joint (c) in the set $C \in \{c_x, c_y, c_z\}$. $\hat{Y}$ denotes the predicted label on a performance component.

This rule simply checks the maximum position of a wrist joint, $p^{\max}(\text{wr}, c_y)$, related to that of a spine shoulder joint, $p^{\max}(\text{spsh}, c_y)$, in the y-coordinate to roughly estimate whether a patient achieves the target position of Exercise 1. For the prediction with multiple rules, we apply a majority voting algorithm and do not apply any tie-breaking method given an odd number of rules.

The score of being correct on each performance component using the RB model (P_{RB}) is computed with the following equation:

$$P_{\text{RB}} = \frac{1}{|\mathbb{R}|} \sum_{r \in R} \min\left(\frac{x_r}{\tau_r}, 1\right) \tag{2}$$

where x_r indicates the feature value of a rule r from a trial (e.g., $p^{\max}(wr, c_y)$ for the example above), τ_r describes the threshold value of a rule r (e.g., $p^{\max}(\text{spsh}, c_y)$ for the example above). $\mathbb{R}$ describes the set of rules elicited from the therapists. min function is applied so that this equation assigns a value of 1 if the feature value of a rule exceeds the threshold of that rule. Otherwise, the equation normalizes the feature value of a rule with the threshold of a rule to compute the score of being correct.

In addition, as the initial threshold values of rules are generic, our approach can further tune a rule-based (RB) model with held-out user's unaffected motions to update its threshold values on each patient (Fig. 1a). For the computation of personalized threshold values, we utilize the held-out user's unaffected motions to learn a Gaussian probability density function $f(x_r) \sim N(\mu_r, \sigma_r^2)$. Specifically, when a patient first interacts with the system and there is no prestored patient's unaffected data, the system will inform the patient to perform an exercise with the patient's unaffected side. When the system has the patient's unaffected data, it will process to extract the feature value of a rule (x_r). We then utilized the maximum likelihood estimate (MLE) (Gopinath

1998) to estimate the parameters of a Gaussian probability density function, μ_r and σ_r as the mean and standard deviation of x_r, respectively. We then update the threshold value for a rule r with either $2\sigma_s$ or $3\sigma_s$ (i.e., $\tau_r \in [\mu_r + 2\sigma_r, \mu_r + 3\sigma_r]$).

4.4 Hybrid model

A hybrid model (HM) applies a weighted average, ensemble technique (Baltrušaitis et al. 2019) to combine machine learning (ML) and rule-based models to assess patients' quality of motion (Lee et al. 2020). For the prediction on the quality of motion, the HM computes the weighted average of prediction scores from two models, in which the contribution weight of each model is the performance of a model (i.e., the F1-score of each model in the range of [0, 1]). Given training data, this weight can be precomputed and updated as the system collects additional data. The equation of computing the score of being correct using the HM, P_{HM}, is as follows:

$$P_{\mathrm{HM}} = \frac{\rho_{\mathrm{ML}}}{\rho_{\mathrm{ML}} + \rho_{\mathrm{RB}}} P_{\mathrm{ML}} + \frac{\rho_{\mathrm{RB}}}{\rho_{\mathrm{ML}} + \rho_{\mathrm{RB}}} P_{\mathrm{RB}} \tag{3}$$

where P_{ML} and P_{RB} indicate the scores of the machine learning (ML) and rule-based (RB) models, and ρ_{ML} and ρ_{RB} describe the weights, F1-scores of ML and RB models.

4.5 Ensemble voting method for frame-level assessment

Our approach can detect a compensation motion at the frame level in real time so that a social robot exercise coaching system can provide a patient with corrective feedback when an erroneous motion has occurred. However, such a frame-level assessment, identifying the exact boundaries of a compensation motion, is challenging (Hasan and Roy-Chowdhury 2014). Thus, our approach applies an ensemble voting method (Dietterich 2000) that utilizes predictions on multiple consecutive V_f frames for a more robust frame-level assessment. The process of this method consists of (1) initial, continuous frame-level predictions and (2) the computation of a score to determine a winning prediction.

Let us denote $h(f_t)$ the predicted frame-level assessment at frame t with an assessment model h (e.g., a machine learning model, a rule-based model, or a hybrid model) and a feature vector f_t. The first process of an initial frame-level prediction runs continuously with an assessment model to generate predicted frame-level assessment $h(f_t)$ at each frame t. When V_f number of initial frame-level predictions are available, our method computes a score of detecting a compensation motion at frame t over all label classes $Y \in \mathcal{Y}$. Then, the winning prediction at frame t is selected as follows:

$$\hat{Y}_t = \underset{Y \in \mathcal{Y}}{\arg\max} \sum_{f_t \in \bar{F}} \delta(h(f_t), Y) \tag{4}$$

where $\bar{F}$ indicates a set of accumulated V_f feature vectors until t frame and $\delta(h(f_t), Y)$ assigns 1 if $h(f_t) = Y$ and 0 otherwise. The δ function is to count the predicted assessment of Y with the predictions from V_f frames. $\hat{Y}_t$ indicates the predicted frame-level assessment at t frame on a compensation motion with the largest number of the predictions, votes from V_f frames. In case of having tied votes, our method assigns $\hat{Y}_t$ with the latest prediction $h(f_t)$. By leveraging votes from past $V_f - 1$ frames to the current t frame, our approach can support a more robust frame-level assessment.

4.6 Interface of a socially assistive robot

Based on our findings from the interviews with therapists (Sect. 3.4), we designed and developed a state machine to enable interactions of our social robot exercise coach system with users as a therapist. This state machine (Fig. 3) includes ten states: *'Greeting/Briefing'*, *'Demonstration'*, *'Initial'*, *'Movement'*, *'Terminate'*, *'Feedback'*, *'Notify'*, *'Encourage'*, *'Correction'*, and *'Wrap-up'*. Depending on the user inputs (e.g., clicking a button to start a system) and the results from our motion analysis component, the state machine will transit to a corresponding state and generate audio and visual feedback and control the behaviors of our social robot exercise coaching system (e.g., gestures).

In the *'Greeting/Briefing'* state, our robotic exercise coaching system will summarize the main goal of a rehabilitation session as specified by a therapist. The system will show the video of a prescribed motion in the *'Demonstration'* state if a new exercise is prescribed and a user requests it. In the *'Initial'* state, the system will prompt whether a user is ready to start an exercise. Once a user confirms to start performing an exercise, the system will transit to the *'Movement'* state and alert that the system starts monitoring in the *'Notify'* state. When a user performs an exercise, the system will provide various types of feedback in the *'Feedback'* state. For instance, if the system detects any compensated motion in real time, the system will provide a user corrective feedback on which unnecessary joints are involved in the *'Correction'* state. Once a user completes an exercise, the state machine of our system transits to the *'Terminate'* state, in which it will summarize the predicted assessment on the quality of motion in the *'Notify'* state and provides *'Encouragement'*. When a user completes all prescribed exercises or requests to finish a session, the system will summarize what a user achieves in the session and remind the next session in the *'Wrap-Up'* state.

For a social assistive robot, we used an NAO robot (SoftBank Robotics Europe, France) that supports competitive hardware capabilities and a user-friendly software development environment with cost reduction (Gouaillier et al. 2008). We utilized the NAO SDK and Choregraphe software (Pot et al. 2009) to program the gestures of the NAO and Google Text To Speech (TTS) APIs (Këpuska and Bohouta 2017) to generate audio outputs from a socially assistive robot exercise coaching system. For communicating the results of motion analysis, we implemented client/server applications with socket programming in Python.

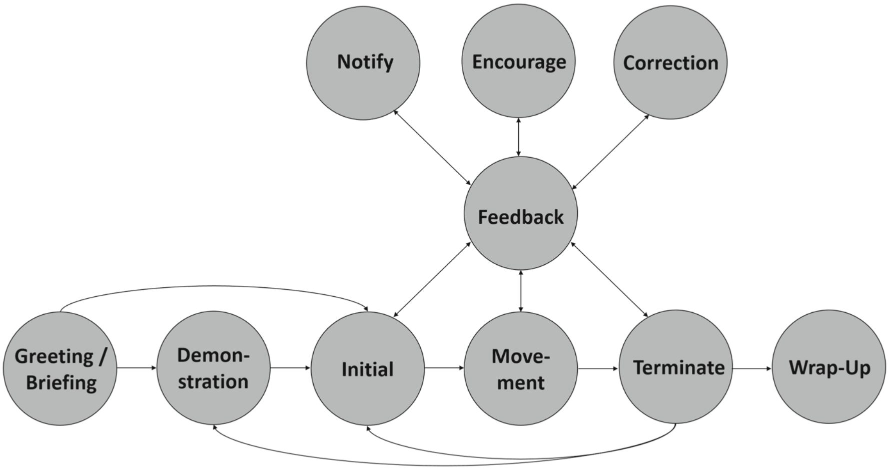

Fig. 3 The state machine of an interactive, social robot exercise coaching system: The system will provide various types of social interactions, such as audio and visual feedback to a user and control the gestures of the social robot based on each state (e.g., greeting, demonstration, feedback on exercise, and wrapping up a session)

5 Experiments

5.1 Dataset of three upper-limb exercises

To evaluate the feasibility of our approach, this work utilizes the dataset of three exercises from 15 post-stroke subjects using a Kinect v2 sensor (Microsoft, Redmond, USA) (Lee et al. 2019). Fifteen post-stroke patients (2 females) with diverse functional abilities from mild to severe impairment (37 ± 21 out of 66 Fugl Meyer Scores (Sanford et al. 1993)) performed 10 repetitions of each exercise with both affected and unaffected sides. During the data collection, a sensor was located at a height of 0.72m above the floor and 2.5m away from a subject and recorded the trajectory of joints and video frames at 30 Hz. The starting and ending frames of exercise movements were manually annotated.

Two therapists (mean (M) = 5.0, standard deviation (SD) = 1.0 years of experience in stroke rehabilitation) annotated the dataset to implement our approach and compute the experts' agreement level. They individually watched the recorded videos of patients' exercise movements and annotated the performance components of the exercise motion dataset. For the frame-level annotation of the *'Compensation'* performance component, two expert annotators reviewed the images that were extracted from the recorded videos with the corresponding sampling frequency using the FFmpeg tool (Developers 2016). The annotations of experts were compared to measure the experts' agreement on F1-scores (i.e., *'Experts' Agreement'* in Appendix Table 4, with the Cohen's kappa of 0.65). We utilized the annotation of an expert, who evaluated the functional abilities of patients with the Fugl–Meyer assessment and had more experience as the ground truth.

The collected dataset was divided into *'Training'* and *'User'* data as follows:

- *'Training Data'* (Fig. 1a) is composed of 140 unaffected motions and 140 affected motions from 14 post-stroke subjects to train a machine learning (ML) model.
- *'User Data'* (Fig. 1a) includes 10 unaffected motions and 10 affected motions of a testing post-stroke subject.

5.2 Quantitative in-laboratory system evaluation

We applied Leave-One-Subject-Out (LOSO) cross-validation on post-stroke patients to evaluate our approach. A machine learning model (ML) was trained with data from all subjects except one testing post-stroke survivor. An initial rule-based (RB) model was developed from the interviews with therapists. A hybrid model applies a weighted average to integrate a trained, outperforming ML model with a rule-based model. All models (e.g., rule-based, machine learning, hybrid) were tested with affected motions of the left-out post-stroke patient. This process was repeated over all post-stroke survivors to evaluate the performance of a model. In addition, we analyzed the effect of tuning a model with held-out unaffected motions of the left-out post-stroke survivor. For a feedforward neural network model, we applied the common transfer learning technique (Zhuang et al. 2020) that fine-tunes a pretrained model with the patient's unaffected motions to implement the tuned feedforward neural network (Tuned ML-NN). We then compared the performance of the Tuned ML-NN with that of the HM-Tuned to evaluate the value of our interactive HM for a personalized assessment. We also explored different numbers of multiple consecutive V_f frames on our ensemble voting method for frame-level assessment (i.e., $V_f = 1, \ldots, 30$). For the performance metric, this work utilized an F1-score that computes the harmonic mean of precision and recall for a more realistic measure of a model.

5.3 Real-world system evaluation

After developing our system, we conducted a real-world experiment to evaluate the potential of our system with healthy participants.

As we had difficulty with running a study with post-stroke patients due to COVID-19, we aimed to conduct a pilot evaluation study to receive early feedback on our system before conducting user studies with post-stroke survivors. For this real-world evaluation, we recruited 10 healthy participants. In each session, the researcher gave an introduction to the study and instructed a compensatory motion of a post-stroke survivor by showing a video and an image of a post-stroke survivor. When a participant became familiar with a compensatory motion, the researcher instructed the participant to perform six repetitions of an exercise: one trial of a correct *'ROM'* and no *'Compensation'*, one trial of an incorrect *'ROM'* and no *'Compensation'*, two trials of a correct *'ROM'* and acted-out *'Compensation'*, and two trials of an incorrect *'ROM'* and acted-out *'Compensation'*. While performing an exercise, our system automatically monitored the participants' exercises and provided real-time feedback through audio, visualization, and robot gestures (Fig. 1b). All sessions were video-recorded for further analysis (e.g., collecting the ground truth). After completing the exercise trials, each participant filled out the following usability questionnaires (Fasola and Matarić 2013) of our system on a 7-point scale and provide any suggestions for improvement:

- Usefulness: *“The system provides a useful, valuable, rich feedback”*
- Intelligence: *“The system is intelligent and competent”*
- Trustfulness: *“The system is trustful”*
- Social Attraction: *“The system is friendly and pleasant. I could have an enjoyable and motivating interaction”*
- Usage Intention: *“I would use the system in future or recommend the system as an exercise partner”*

As it was difficult to instruct and act out post-stroke survivors’ acted-out non-smooth motions, we excluded to act-out *‘Smoothness’* component during the study. A researcher, who facilitated the evaluation experiment, only manually indicated a starting cue to start performing an exercise trial. All other functionalities of our system (Fig. 1b) were operated autonomously during the study. The protocols of this user study were reviewed and approved by the institutional review board.

6 Results

6.1 In-laboratory system performance

Figure 4 summarizes the performances of models, which measure an agreement with ground-truth labels by computing average F1-scores on performance components of three exercises. For machine learning (ML) models, we explored a non-interactive, feedforward neural network (ML-NN), building upon the results from Lee et al. (2020). In addition, we presented the results of a tuned, feedforward neural network (Tuned ML-NN). The parameters of ML-NN models (i.e., hidden layers/units and learning rates of feedforward neural networks) that achieved the best F1-score during leave-one-subject-out (LOSO) cross-validation are summarized in Appendix Table 3.

In addition, we present the performance of the initial, non-interactive rule-based model (Non-interactive RB) from the interviews with therapists and that of the interact fine-tuned rule-based model (RB-tuned) after leveraging the held-out user’s unaffected motions to tune threshold values for a personalized assessment. The parameters of rule-based models (i.e., the range of the threshold values with 2σ or 3σ) are selected to achieve the best F1-score during validation: 3σ is utilized over three performance components of three exercises except for the *‘ROM’* and *‘Smoothness’* of both Exercises 1 and 2.

For hybrid models (HMs), we describe the performance of the initial, non-interactive hybrid model (Non-Interactive HM) that integrates the feedforward neural network (ML-NN) with the non-interactive rule-based model (Non-Interactive RB) and that of the interactive, tuned hybrid model (HM-tuned) that combines the ML-NN with the interactive, tuned rule-based model (RB-Tuned).

For machine learning (ML) models, neural networks (ML-NN) achieve a good agreement level with ground-truth annotations (i.e., 0.7899 average F1-score over all exercises), which is equally good with experts’ agreement. However, the initial, non-interactive rule-based model (Non-Interactive RB) achieves the lowest performance: 0.5827 average F1-score over all exercises. According to the further analysis of the non-interactive, rule-based model, we found that such low performance occurred, because elicited rules from therapists are generic and not tuned for individuals with different physical conditions (Lee et al. 2020). For instance, one rule of assessing the *'Compensation'* performance component is to check whether the x-coordinate of a shoulder joint is located more than the 15% of the initial position. We found that even if affected motions of some patients were annotated as normal and did not involve compensated shoulder movements, the shoulder joint of those motions was located around 20% of the initial positions, which was greater than a generic threshold value and was misclassified as compensated motions. This indicates the importance of generating personalized rules for patients with various physical characteristics and functional abilities.

The initial, non-interactive hybrid model (Non-Interactive HM) achieved a 0.7447 average F1-score over all exercises. As the initial, non-interactive rule-based model (Non-Interactive RB) had limited performance, the non-interactive HM that integrates the ML model with neural networks (ML-NN) and the non-Interactive RB led to slightly lower performance than that of the ML-NN (i.e., 0.7899 average F1-score over all exercises). However, the non-interactive HM still achieved comparable performance to the experts' agreement.

To evaluate the feasibility of tuning a model for personalized assessment, we updated the threshold values of a rule-based model with held-out patient's unaffected motions (as described in Sect. 4.3) and implemented the interactive, tuned rule-based model (RB-Tuned) and interactive, tuned hybrid model (HM-Tune) that integrates the ML-NN model with the interactive, RB-Tuned model. In addition, we implemented the tuned neural network model (Tuned ML-NN) that fine-tunes a neural network model (ML-NN) with the patient's unaffected motions using the common transfer learning technique (Zhuang et al. 2020). We then compared the performance of the Tuned ML-NN with that of the interactive, HM-Tuned to evaluate the value of our interactive HM for personalized assessments.

Both RB-Tuned and HM-Tuned models significantly improved their performance to replicate the therapist's assessment ($p < 0.01$ using the paired t-tests over three performance components of three exercises). Specifically, the RB model significantly improved its performance around 37% from 0.5821 to 0.7957 average F1-scores over all exercises ($p < 0.01$). In addition, the hybrid model (HM) also significantly improved its performance around 11% from 0.7447 to 0.8235 average F1-scores over all exercises ($p < 0.01$) and outperformed other approaches. The performance of the tuned hybrid model (HM-tuned) was better than those of the machine learning model with neural networks (ML-NN), the Tuned ML-NN, and the RB-Tuned (i.e., 4%, 10%, and 3% improvement, respectively, without statistical significance). Unlike our RB-Tuned and HM-Tuned models that improved their performance, the Tuned ML-NN performed worse 5% from 0.7899 to 0.7427 average F1-scores after tuning with the patient's unaffected motions.

To analyze the effect of our ensemble voting method for frame-level assessment, we utilized the ML-NN, RB-Tuned, and HM-Tuned models and plotted their average performance of detecting frame-level compensation on the head, spine, shoulder joints

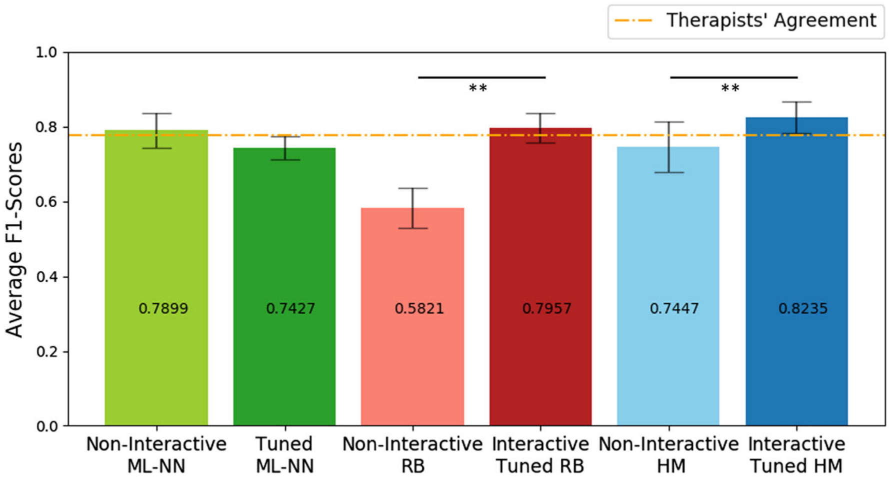

Fig. 4 Comparison of model performance without/with tuning with user data: Both the rule-based (RB) model and the hybrid model (HM) significantly improved their performance to replicate the therapist's assessment while tuning with patient's unaffected motions. The RB model significantly improved its performance by 37% from 0.5821 to 0.7957 average F1-score ($p < 0.01$ using paired t-tests) and the HM improved its performance by 11% from 0.7447 to 0.8235 average F1-score over three exercises ($p < 0.01$ using paired t-tests). In contrast to the RB and HM models, the Tuned ML-NN performed worse than the ML-NN after tuning with the patient's unaffected motions

over three exercises with various numbers of consecutive frames ($V_f = 1, \ldots, 30$). In Fig. 5, all three models (i.e., ML-NN, RB-Tuned, HM-Tuned) improved their performance while leveraging prediction from multiple frames and achieved their best performance with $V_f = 29$. When we compared the performance of a model without and with an ensemble voting method ($V_f = 1$ and $V_f = 29$), the ML-NN model improved its performance from 0.7723 ($V_f = 1$) to 0.7803 ($V_f = 29$) average F1-score ($p < 0.01$ using the paired t-tests over three compensations of three exercises); the RB-Tuned model improved its performance from 0.7655 ($V_f = 1$) to 0.7816 ($V_f = 29$) average F1-score ($p < 0.01$); the HM-Tuned model improved its performance from 0.7975 ($V_f = 1$) to 0.8070 ($V_f = 29$) average F1-score ($p < 0.01$).

6.2 Real-world system performance

Figure 6 summarizes the performance of non-interactive and interactive hybrid models during the in-laboratory and real-world studies. Our results showed that non-interactive models of the real-world study led to lower average performance compared to the models of the in-laboratory study along with the performance degradation of 22% from 0.84 F1-score to 0.65 F1-score. Also, interactive models of the real-world study led to lower average performance than the models of the in-laboratory study. However, our interactive models led to a performance degradation of 5% from 0.86 F1-score to 0.81 F1-score, which is less than that of non-interactive models. Our system could still adapt to new participants with diverse physical characteristics and achieved performance that is comparable with experts' agreement.

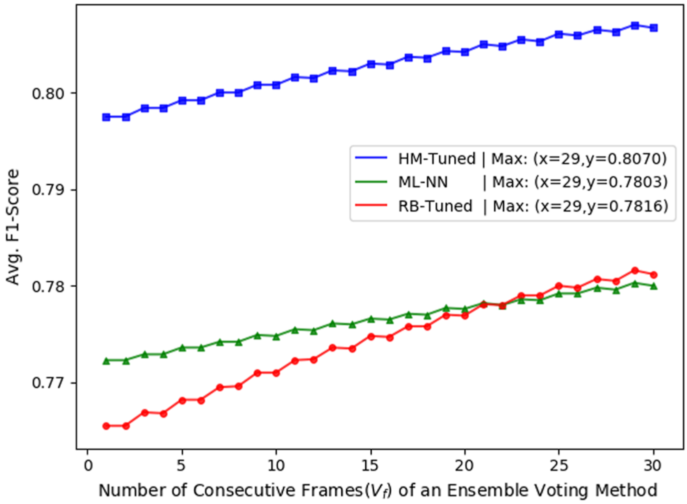

Fig. 5 Performance of frame-level assessment with different numbers of consecutive frames (V_f) using the tuned rule-based model (RB-Tuned), machine learning model with neural networks (ML-NN), tuned hybrid model (HM-Tuned)

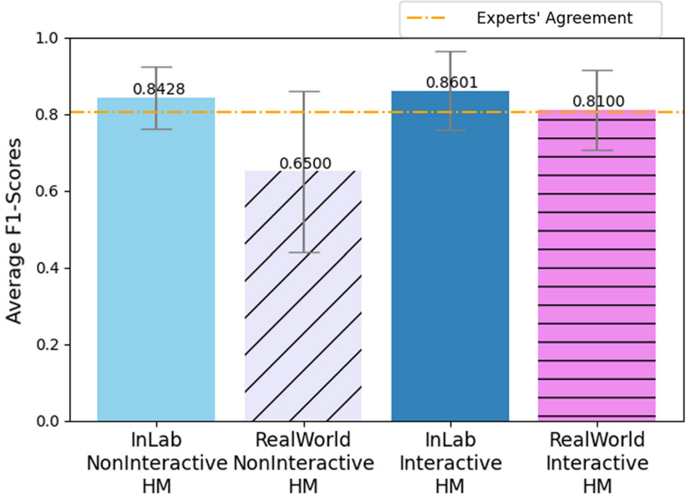

Fig. 6 Comparison of performance of non-interactive and interactive hybrid models for quantitative rehabilitation assessment during in-laboratory and real-world studies: Although our interactive hybrid model achieved slightly lower performance during a real-world study, our interactive model can still adapt to a new participant and achieve comparable performance to experts' agreement unlike non-interactive models. Compared to the non-interactive HM model, our interactive HM model had much lower performance degradation from the in-laboratory study to the real-world study

In addition, Fig. 7 describes the histogram of usability responses from participants in the real-world study. Overall, participants in the real-world study expressed positive opinions on our interactive robotic exercise coach. They appreciated our system that is *"useful to observe my [their] body alignment and positions"* (P5) during exercises

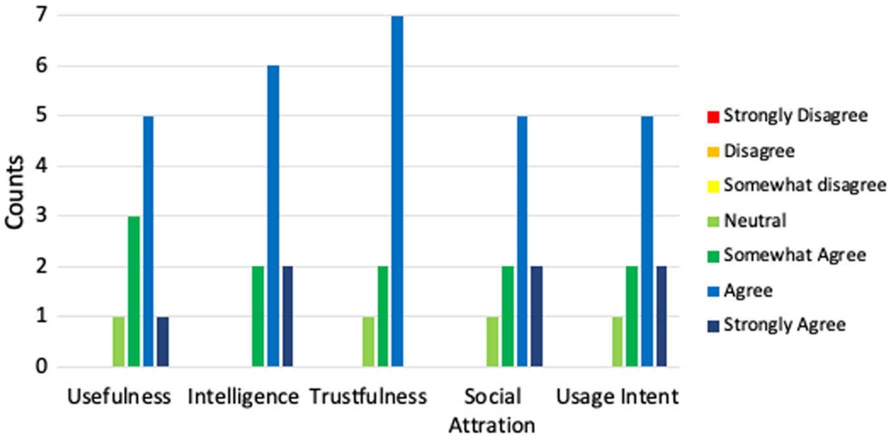

Fig. 7 Usability survey responses from participants during the real-world evaluation study: most participants were positive to use our system which is considered as socially attractive and intelligent by providing useful and trustful information

and provide feedback on *"how well I [a participant] did it [an exercise]"* (P6). Even if our system provides incorrect assessment, participants considered that feature-specific feedback from our system assisted them to better determine whether to trust the feedback of a system: *"I find the system trustful because the feedback made me clear how I should improve in a certain way"* (P9). Participants also enjoyed having an exercise with a robot, but one participant considered that it would be better if the robot can have *"a more friendly face or appearance"*(P3). Overall, participants were positive to use or recommend our system as an exercise coach even with some limitations.

7 Discussion

In this work, we study and discuss how a social robot exercise coaching system can be designed and developed to generate personalized corrective feedback along with the in-laboratory and real-world evaluation studies.

For generating personalized corrective feedback, we compared various existing approaches with our proposed hybrid model and evaluated the effect of an ensemble voting method for real-time, frame-level assessment (Lee et al. 2020). Among various approaches, the machine learning model with neural networks (ML-NN), the tuned rule-based model (RB-tuned), and the initial, non-interactive and the interactive, tuned hybrid models (Non-Interactive HM and Interactive, Tuned HM) have equally good performance with expert's agreement from the paired *t*-tests over three performance components of three exercises. In addition, all models with an ensemble voting method can leverage predictions from multiple consecutive frames to improve their frame-level assessment and inform a user when an erroneous motion has occurred.

As a rule-based (RB) model does not require the data collection process, an RB model could be considered as a natural starting point to develop a social robot exercise coaching system that can assess the quality of motion and generate corrective feedback on patient's exercises (Matarić et al. 2007; Lee et al. 2020). However, an RB model with generic threshold values (e.g., RB-Init and (Fasola and Matarić 2013; Görer et al. 2017; Lee et al. 2020; Guneysu and Arnrich 2017; Tanguy et al. 2016)) does not perform well to evaluate exercises of patients with various physical conditions. Thus, it is important to have an interactive approach that can tune an RB model with individual's held-out unaffected motions to derive personalized threshold values for assessment and corrective feedback.

When a social robot exercise coaching system is deployed and annotated data is collected, a machine learning (ML) model (e.g., neural networks) can be trained to extract new insights for assessing exercises from data. However, we do not recommend simply replacing a rule-based (RB) model with an ML model using a complex algorithm that operates as a black box model. For instance, given a patient's affected motion that is incorrectly performed with compensation, an ML model with neural networks can just notify whether the compensation has occurred or not without any explanations on the outputs of the model (Rudin and Radin 2019). In contrast, our interactive hybrid model can predict assessment with improved performance, but also identify which feature has been violated with a rule-based model: the violation on the head in the z-axis and the shoulder in the y-axis for Fig. 2b. Such feature-level analysis can be realized in the following personalized corrective feedback: *"Keep your head straight and do not raise your shoulder"* (Lee et al. 2020, 2022). We found that participants in our real-world study appreciated the potential of our system to make them have trustful interactions with it. Thus, after data collection, a hybrid model is recommended to accommodate new generic insights from data and support a transparent and personalized interaction between a robot and a user.

When it comes to the evaluation of the system performance, we found that our in-laboratory study through leave-one-subject-out (LOSO) cross-validation has a slightly over-promising performance than our real-world study. However, the performance difference is not statistically significant using the t-test. Thus, we considered that the LOSO cross-validation has the potential to provide the estimated system performance in practice, which still needs to be carefully analyzed further (Rao et al. 2008).

During our real-world study, we found that our interactive robotic exercise coach has the potential to adapt to a new user and automatically monitor participants' exercises and provide personalized corrective feedback. However, our system implementation still requires manual input from a researcher to indicate the starting time of the user's motion. For creating a fully autonomous system, the exploration of techniques for motion segmentation (Lin and Kulić 2013) is necessary. In addition, we have only conducted the pilot evaluation with healthy participants, who acted out post-stroke survivor's motions. In-person user studies with post-stroke survivors are required to better understand the feasibility of our system in practice. As post-stroke survivors might perform incorrect motions that might exacerbate their conditions, it is also important to explore a way to adapt a rehabilitation session and program (Lee et al. 2022) beyond personalized feedback that has been studied in this work.

8 Conclusion

In this paper, we contributed to the designs, development, and evaluation of an interactive approach with an ensemble voting method for a social robot exercise coach system in the context of physical stroke rehabilitation therapy. This system integrates a machine learning model with an interactive and interpretable rule-based model and tunes with patient's data for real-time, personalized corrective feedback on patient's exercises. Through in-laboratory and real-world experiments, this work shows that our interactive hybrid model can adapt to a new user and achieve better performance to replicate an expert's assessment and feedback on unobserved data of new users, but also support transparent and personalized interaction of a robotic exercise coaching system. In addition, this work discusses the potential benefits and limitations of our system to support post-stroke survivor's rehabilitation sessions.

Acknowledgements The authors thank all the participants in this study for their time and valuable inputs. This work was partially supported by the National Science Foundation (NSF) under grant number CNS-1518865. Additional support was provided by the IntelligentCare project (LISBOA-01-0247-FEDER-045948), the FCT LARSyS funding 2020-2023 (UIDB/50009/2020), the FCT project HAVATAR (PTDC/EEI-ROB/1155/2020), and the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

Declarations

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Appendix A

See Tables 3, 4 and 5.

Table 3 Parameters of machine learning models

	Hidden layers and units/learning rate		
	ROM	Smoothness	Comp
E1	– NN: (256, 256, 256)/0.005	– NN: (16)/0.0001	– NN: (512, 512, 512)/0.005
E2	– NN: (32, 32, 32)/0.01	– NN: (32)/0.0001	– NN: (256, 256)/0.0001
E3	– NN: (16)/0.005	– NN: (128)/0.0001	– NN: (256, 256, 256)/0.1

Table 4 Performances (avg. ± std. of F1-scores) of machine learning (ML) models, rule-based (RB) models, hybrid models (HMs), and experts' agreement

Algorithm	Exercise 1	Exercise 2	Exercise 3	Overall
ML-NN	0.8428 ± 0.0809	0.7549 ± 0.1026	0.7720 ± 0.0433	0.7899 ± 0.0466
Tuned ML-NN	0.7707 ± 0.1093	0.7105 ± 0.1308	0.7470 ± 0.0381	0.7427 ± 0.0303
RB-Init‡	0.6148 ± 0.2086	0.6707 ± 0.1758	0.4626 ± 0.2102	0.5827 ± 0.0541
RB-Tuned	0.8317 ± 0.0784	0.8009 ± 0.1238	0.7543 ± 0.0248	0.7957 ± 0.0390
HM-Init‡	0.8069 ± 0.0946	0.7060 ± 0.1318	0.7212 ± 0.0851	0.7447 ± 0.0679
HM-Tuned	0.8601 ± 0.1030	0.7769 ± 0.1317	0.8334 ± 0.1142	**0.8235 ± 0.0425**
Experts' Agreement	0.7908 ± 0.2146	0.8222 ± 0.1534	0.7196 ± 0.1754	0.7775 ± 0.0526

‡ indicates HM-Tuned performs statistically better than the compared method (pairwise *t*-tests at 99% significance level)

The highest performance of the model is boldfaced

Table 5 List of independent rules to assess the quality of motion from therapists

Performance components	Rules
Range of motion (ROM)	A wrist joint should be located above a spine-shoulder joint near a head joint for exercise 1
	A wrist joint should be located higher than a shoulder joint for exercise 2
	A wrist joint should be located further than hip near a knee for exercise 3
Smoothness	A wrist joint should be smoothly coordinated in the *x*-axis during 80% of the motion
	(Zero-crossing ratio of a wrist acceleration in the *x*-axis is within 20%)
	A wrist joint should be smoothly coordinated in the *y*-axis during 80% of the motion
	(Zero-crossing ratio of a wrist acceleration in the *y*-axis is within 20%)
	A wrist joint should be smoothly coordinated in the *z*-axis during 80% of the motion
	(Zero-crossing ratio of a wrist acceleration in the *z*-axis is within 20%)
Compensation	A head joint should not be located more/less than 15% of an initial head position in the *x*-axis
	A head joint should not be located above/below 15% of an initial head position in the *y*-axis
	A head joint should not be located more/less than 15% of an initial head position in the *z*-axis
	A spine joint should not be located more/less than 15% of an initial spine position in the *x*-axis
	A spine joint should not be located above/below 15% of an initial spine position in the *y*-axis

Table 5 continued

Performance components	Rules
	A spine joint should not be located more/less than 15% of an initial spine position in the *z*-axis
	A shoulder joint should not be located more/less than 15% of an initial shoulder position in the *x*-axis
	A shoulder joint should not be located above/below 15% of an initial shoulder position in the *y*-axis
	A shoulder joint should not be located more/less than 15% of an initial shoulder position in the *z*-axis

References

Azenkot, S., Feng, C., Cakmak, M.: Enabling building service robots to guide blind people a participatory design approach. In: 2016 11th ACM/IEEE International Conference on Human–Robot Interaction (HRI), pp. 3–10. IEEE (2016)

Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2019)

Beer, J.M., Smarr, C.A., Chen, T.L., Prakash, A., Mitzner, T.L., Kemp, C.C., Rogers, W.A.: The domesticated robot: design guidelines for assisting older adults to age in place. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human–Robot Interaction, pp. 335–342 (2012)

Cirstea, M., Levin, M.F.: Compensatory strategies for reaching in stroke. Brain **123**(5), 940–953 (2000)

Dall, T.M., Gallo, P.D., Chakrabarti, R., West, T., Semilla, A.P., Storm, M.V.: An aging population and growing disease burden will require alarge and specialized health care workforce by 2025. Health Aff. **32**(11), 2013–2020 (2013)

Developers, F.: ffmpeg tool. http://ffmpeg.org (2016)

Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems, pp. 1–15. Springer (2000)

Fasola, J., Matarić, M.J.: A socially assistive robot exercise coach for the elderly. J. Hum. Robot Interact. **2**(2), 3–32 (2013)

Feigin, V.L., Norrving, B., Mensah, G.A.: Global burden of stroke. Circ. Res. **120**(3), 439–448 (2017)

Feil-Seifer, D., Mataric, M.J.: Defining socially assistive robotics. In: 9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005, pp. 465–468. IEEE (2005)

Feingold Polak, R., Tzedek, S.L.: Social robot for rehabilitation: expert clinicians and post-stroke patients' evaluation following a long-term intervention. In: Proceedings of the 2020 ACM/IEEE International Conference on Human–Robot Interaction, pp. 151–160 (2020)

Gale, N.K., Heath, G., Cameron, E., Rashid, S., Redwood, S.: Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med. Res. Methodol. **13**(1), 1–8 (2013)

Gopinath, R.A.: Maximum likelihood modeling with gaussian distributions for classification. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181), vol. 2, pp. 661–664. IEEE (1998)

Görer, B., Salah, A.A., Akın, H.L.: A robotic fitness coach for the elderly. In: International Joint Conference on Ambient Intelligence, pp. 124–139. Springer (2013)

Görer, B., Salah, A.A., Akın, H.L.: An autonomous robotic exercise tutor for elderly people. Auton. Robot. **41**(3), 657–678 (2017)

Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., Maisonnier, B.: The nao humanoid: a combination of performance and affordability. CoRR arXiv:0807.3223 (2008)

Guneysu, A., Arnrich, B.: Socially assistive child-robot interaction in physical exercise coaching. In: 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 670–675. IEEE (2017)

Hasan, M., Roy-Chowdhury, A.K.: Continuous learning of human activity models using deep nets. In: European Conference on Computer Vision, pp. 705–720. Springer (2014)

Irfan, B., Gomez, N.C., Casas, J., Senft, E., Gutiérrez, L.F., Rincon-Roncancio, M., Munera, M., Belpaeme, T., Cifuentes, C.A.: Using a personalised socially assistive robot for cardiac rehabilitation: a long-term case study. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 124–130. IEEE (2020)

Kåringen, I., Dysvik, E., Furnes, B.: The elderly stroke patient's long-term adherence to physiotherapy home exercises. Adv. Physiother. **13**(4), 145–152 (2011)

Kashi, S., Polak, R.F., Lerner, B., Rokach, L., Levy-Tzedek, S.: A machine-learning model for automatic detection of movement compensations in stroke patients. IEEE Trans. Emerg. Top. Comput. **9**(3), 1234–1247 (2020)

Këpuska, V., Bohouta, G.: Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). Int. J. Eng. Res. Appl. **7**(03), 20–24 (2017)

Lee, M.H., Siewiorek, D.P., Smailagic, A., Bernardino, A., Bermúdez i Badia, S.: 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, UMAP '20, pp. 303–307. ACM (2020)

Lee, M.H., Siewiorek, D.P., Smailagic, A., Bernardino, A., Bermúdez i Badia, S.: Interactive hybrid approach to combine machine and human intelligence for personalized rehabilitation assessment. In: Proceedings of the ACM Conference on Health, Inference, and Learning, pp. 160–169 (2020)

Lee, M.H., Siewiorek, D.P., Smailagic, A., Bernardino, A., Badia, S.B.: Towards personalized interaction and corrective feedback of a socially assistive robot for post-stroke rehabilitation therapy. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1366–1373. IEEE (2020)

Lee, M.H., Siewiorek, D.P., Smailagic, A., Bernardino, A., Badia, S.B.i.: Learning to assess the quality of stroke rehabilitation exercises. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 218–228 (2019)

Lee, M.H., Siewiorek, D.P., Smailagic, A., Bernardino, A., et al.: Designing personalized interaction of a socially assistive robot for stroke rehabilitation therapy. arXiv:2007.06473 (2020)

Lee, M.H., Siewiorek, D.P., Smailagic, A., Bernardino, A., et al.: Enabling AI and robotic coaches for physical rehabilitation therapy: iterative design and evaluation with therapists and post-stroke survivors. Int. J. Soc. Robot. 1–22 (2022)

Lin, J.F.S., Kulić, D.: Online segmentation of human motion for automated rehabilitation exercise analysis. IEEE Trans. Neural Syst. Rehabil. Eng. **22**(1), 168–180 (2013)

Matarić, M., Tapus, A., Winstein, C., Eriksson, J.: Socially assistive robotics for stroke and mild TBI rehabilitation. In: Advanced Technologies in Rehabilitation, pp. 249–262. IOS Press (2009)

Matarić, M.J., Eriksson, J., Feil-Seifer, D.J., Winstein, C.J.: Socially assistive robotics for post-stroke rehabilitation. J. Neuroeng. Rehabil. **4**(1), 5 (2007)

Matarić, M.J., Scassellati, B.: Socially Assistive Robotics. Springer Handbook of Robotics, pp. 1973–1994 (2016)

O'Sullivan, S.B., Schmitz, T.J., Fulk, G.: Physical Rehabilitation. F. A. Davis (2019)

Pot, E., Monceaux, J., Gelin, R., Maisonnier, B.: Choregraphe: a graphical tool for humanoid robot programming. In: RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication, pp. 46–51. IEEE (2009)

Rao, R.B., Fung, G., Rosales, R.: On the dangers of cross-validation. an experimental evaluation. In: Proceedings of the 2008 SIAM International Conference on Data Mining, pp. 588–596. SIAM (2008)

Riek, L.D.: Healthcare robotics. Commun. ACM **60**(11), 68–78 (2017)

Rudin, C., Radin, J.: Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition (2019)

Sanford, J., Moreland, J., Swanson, L.R., Stratford, P.W., Gowland, C.: Reliability of the Fugl–Meyer assessment for testing motor performance in patients following stroke. Phys. Ther. **73**(7), 447–454 (1993)

Schneider, S., Kummert, F.: Comparing robot and human guided personalization: adaptive exercise robots are perceived as more competent and trustworthy. Int. J. Soc. Robot. **13**(2), 169–185 (2021)

Tanguy, P., Rémy-Néris, O., et al.: Computational architecture of a robot coach for physical exercises in kinaesthetic rehabilitation. In: 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 1138–1143. IEEE (2016)

Tapus, A., Maja, M., Scassellatti, B.: The grand challenges in socially assistive robotics (2007)

Tapus, A., Mataric, M.J.: Towards socially assistive robotics. J. Robot. Soc. Jpn. **24**(5), 576–578 (2006)

Tapus, A., Tapus, C., Mataric, M.J.: The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia. In: 2009 IEEE International Conference on Rehabilitation Robotics, pp. 924–929. IEEE (2009)

Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data **3**(1), 1–40 (2016)

Winkle, K., Caleb-Solly, P., Turton, A., Bremner, P.: Social robots for engagement in rehabilitative therapies: Design implications from a study with therapists. In: Proceedings of the 2018 ACM/IEEE International Conference on Human–Robot Interaction, pp. 289–297. ACM (2018)

Wolf, S.L., Catlin, P.A., Ellis, M., Archer, A.L., Morgan, B., Piacentino, A.: Assessing wolf motor function test as outcome measure for research in patients after stroke. Stroke **32**(7), 1635–1639 (2001)

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proc. IEEE **109**(1), 43–76 (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.