2020 29th IEEE International Conference on
Robot and Human Interactive Communication (RO-MAN)
August 31 - September 4, 2020 - Virtual Conference

Towards Personalized Interaction and Corrective Feedback of a Socially Assistive Robot for Post-Stroke Rehabilitation Therapy

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

Abstract— **A robotic exercise coaching system requires the capability of automatically assessing a patient's exercise to interact with a patient and generate corrective feedback. However, even if patients have various physical conditions, most prior work on robotic exercise coaching systems has utilized generic, pre-defined feedback.**

This paper presents an interactive approach that combines machine learning and rule-based models to automatically assess a patient's rehabilitation exercise and tunes with patient's data to generate personalized corrective feedback. To generate feedback when an erroneous motion occurs, our approach applies an ensemble voting method that leverages predictions from multiple frames for frame-level assessment. According to the evaluation with the dataset of three stroke rehabilitation exercises from 15 post-stroke subjects, our interactive approach with an ensemble voting method supports more accurate frame-level assessment ($p < 0.01$), but also can be tuned with held-out user's unaffected motions to significantly improve the performance of assessment from 0.7447 to 0.8235 average F1-scores over all exercises ($p < 0.01$). This paper discusses the value of an interactive approach with an ensemble voting method for personalized interaction of a robotic exercise coaching system.

I. Introduction

Patients with neurological and musculoskeletal problems (e.g. stroke) require early and extensive physical therapy sessions with task-oriented exercises for months to regain their functional ability [1]. During a session, a therapist monitors and assesses patient's exercises to provide corrective feedback. However, patients can receive the limited amount of those supervised sessions due to the shortage of therapists and the costs [2]. Instead, in-home rehabilitation regimens are often prescribed. During in-home rehabilitation regimens, patients might become confused whether they correctly perform exercises and lose their motivation without any supervision[2].

Recent advances in computing and artificial intelligence empowers a robot with various autonomous capabilities to understand and interact with the world [3]. Researchers have explored the possibility of supplementing health services with advanced computing and socially assistive robotics [4]. For instance, researchers have envisioned that a robotic exercise coaching system can be integrated into a rehabilitation process by automatically monitoring patient's exercises and providing motivational and corrective feedback until the patient's next visits to a therapist [4], [5]. Prior work on robotic exercise coaching systems has demonstrated that elderly or post-stroke subjects can successfully exercise and stay engaged with a robot over sessions [6], [7]. However, in spite of this potential of a robot to monitor and guide exercises, prior work is limited to provide generic, pre-defined corrective feedback on patient's exercise performance (e.g. checking angular difference with the pre-specified motion [6], [7]). It is still challenging to empower a robotic exercise coaching system to generate tailored corrective feedback on an individual patient's motion [7].

In this paper, we focus on improving an approach to automatically assess exercise performance to generate personalized corrective feedback that can be the basis of more intelligent and natural interaction between a robotic exercise coach system and patients afterwards. Specifically, this paper presents an interactive approach that integrates the benefits of machine learning (ML) and rule-based (RB) models to assess the performance of exercises for personalized post-stroke therapy (Figure 1a). As a ML model is able to automatically learn insights on a large amount of data, our approach leverages this ML model to achieve generic assessment on patient's quality of motion. In addition, our approach utilizes a RB model that is flexible to address an edge case of patients with diverse characteristics and interpretable to generate personalized feedback. Our approach applies a weighted average ensemble technique [8] to derive a hybrid model (HM) that accommodates these two perspectives on assessment from both ML and RB models [9]. Moreover, our approach applies an ensemble voting method that utilizes evidence, predictions of multiple consecutive frames for frame-level assessment, and generates patient feedback when an erroneous motion is occurred. Given a new patient, our approach first tunes a RB model with the patient's unaffected motions. Our approach then predicts the quality of motion and generates personalized corrective feedback on patient's affected motions (Figure 1c).

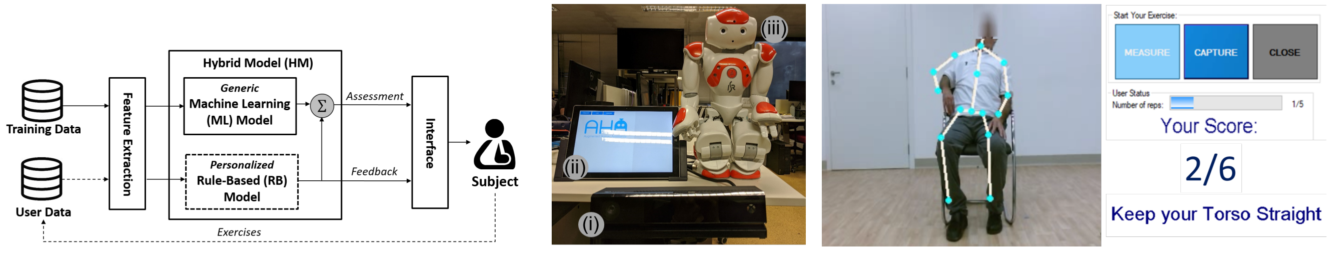
For the development and evaluation of our approach, we utilize the dataset of three upper-limb rehabilitation exercises from 15 post-stroke subjects with the corresponding annotation from two experts. Given this dataset, a machine learning (ML) model with neural network is trained using leave-one-subject-out cross validation. For the initial development of a rule-based (RB) model, we conducted semi-structured interviews with therapists to elicit their knowledge of assessing

[1] Carnegie Mellon University, Pittsburgh, PA 15213, USA {minhunl,dps,asim}@cs.cmu.edu

[2] Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal alex@isr.tecnico.ulisboa.pt

[3] Madeira Interactive Technology Institute, University of Madeira, NOVA-LINCS, Funchal, Portugal sergi.bermudez@m-iti.org

978-1-7281-6075-7/20/$31.00 ©2020 IEEE



(a) (b) (c)

Fig. 1: (a) Flow diagram of an interactive approach of a socially assistive robot for personalized physical therapy. (b) the setup of a system with (i) a Kinect sensor, (ii) a tablet with the visualization interface, and (iii) the NAO robot. (c) An example output of the visualization interface that presents predicted assessment on patient's exercise performance with corrective feedback.

stroke rehabilitation exercises.

After implementing our interactive hybrid model (HM), we empirically evaluate the performance of a model to replicate expert's assessment, and analyze the effect of tuning a model with patient's unaffected motions and a majority voting method for frame-level assessment. Our experimental results show that an initial, generic HM can be tuned into a personalized model with patient's unaffected motions while significantly improving the performance of the HM around 11% from 0.7447 to 0.8235 average F1-scores on three exercises ($p < 0.01$). In addition, all models (i.e. machine learning, rule-based, and hybrid models) with an ensemble voting method improve their performance of frame-level assessment ($p < 0.01$).

This paper makes the following contributions:

- We present approaches to improve the capability of a robot exercise coaching system to automatically assess exercise performance and generate personalized corrective feedback: 1) an interactive approach that combines machine learning and rule-based models and 2) an ensemble voting method for frame-level assessment
- We demonstrate the effect of tuning a model with patient's unaffected motions and an ensemble voting method for frame level assessment, and discuss the value of an interactive approach with an ensemble voting method for personalized, transparent interaction of a robotic exercise coaching system.

II. Related Work

The research of socially assistive robotics has shown great potential to supplement healthcare services through social interaction [10]. For instance, a robotic exercise coaching system can be deployed in a rehabilitation process to automatically monitor rehabilitation exercises and provide subjects feedback without the presence of a therapist [4], [5]. Fasola and Mataric demonstrated that elderly people considered a physically embodied robot more engaging and acceptable as an exercise partner than a virtually embodied agent [6]. Furthermore, Researchers have shown that diverse populations (i.e. post-stroke patients [6], elderly people [7], children [11]) can successfully engage in exercise sessions with a robotic exercise coaching system.

For a robotic exercise coaching system, the capability of automatically assessing a patient's motion is critical to derive a personalized interaction with tailored corrective feedback on patient's exercise performance [7], [12]. However, limited prior work on robotic exercise coaching systems has explored how an automated assessment approach can be developed to generate personalized corrective feedback. For instance, researchers have implemented a method that evaluates the completion of an exercise by computing the difference of a joint angle between user's motion and the pre-defined target motion [6], [7]. Guneysu and Arnrich [11] applied dynamic time warping to compute the statistics of a joint angle and distance measures with a pre-defined motion. Nguyen et al. [13] 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 [11] and [13] support to analyze multiple variables for evaluating an exercise, they still rely on a pre-defined motion or a generic threshold. Prior work with generic threshold-based methods might not be applicable for patients with various characteristics [9]. In addition, there is a lack of evaluation on how well these methods can monitor other complex performance metrics of an exercise (e.g. smoothness or the occurrence of a compensation motion).

For personalized rehabilitation assessment, Lee et al. have showed that an interactive approach can dynamically select features of assessment using reinforcement learning to generate patient-specific analysis [14] as a decision support tool for therapists [15] and a robotic coaching system for patients [12]. However, prior work is limited to provide assessment after completing a motion and does not support frame-level assessment to provide any information on when an erroneous motion has occurred.

In contrast to prior work described above, we present an interactive approach that combines machine learning (ML) and rule-based (RB) models to assess the quality of motion and tune with patient's data to generate personalized corrective feedback of a socially assistive robotic system for physical therapy. Our approach can support complex multivariate analysis on patient's exercises with a machine learning model. In addition, our approach can be easily updated with a RB model to accommodate individual's diverse physical characteristics for personalized assessment and feedback. Instead of tuning a model with therapist's feedback [14], our approach tunes a RB model with held-out patient's data. To generate feedback when an erroneous motion has occurred, our approach supports frame-level assessment of a compensated motion with an ensemble voting method that accommodates predictions on multiple consecutive frames of an exercise motion. Using the annotated dataset of three upper-limb stroke rehabilitation exercises from 15 post-stroke patients, we demonstrate how well our approach can automatically assess the overall quality of motion (e.g. range of motion and smoothness) and detect a compensation motion on head, spine, shoulder joints at frame level. This work contributes to increasing knowledge on techniques to automatically assess exercises for a robotic exercise coaching system.

III. Study for Stroke Rehabilitation

We selected a probe domain as stroke, which is the second leading cause of death and third most common contributor to disability [16]. We had iterative discussion with three therapists ($\mu = 6.33$, $\sigma = 2.05$ years of experience in stroke rehabilitation) to specify the study designs on stroke rehabilitation: exercises and performance components for assessment [17].

A. Three Task-Oriented Upper Limb Exercises

This work utilizes three upper-limb stroke rehabilitation exercises recommended by therapists [17]. For Exercise 1, a subject has to raise the subject's wrist to the mouth as if drinking water. For Exercise 2, a subject has to pretend to touch a light switch on the wall. For Exercise 3, a subject has to extend the subject's elbow in the seated position to practice the usage of a cane.

B. Performance Components

We discussed commonly used stroke assessment tools (i.e. the Wolf Motor Function Test and the Fugl Meyer Assessment [18]) with therapists and specified three common performance components of stroke rehabilitation exercises: *'Range of Motion (ROM)'*, *'Smoothness'*, and *'Compensation'* [17]. 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 compensated movements to achieve a target movement. For instance, a patient might lean the patient's head or trunk to the side and elevate the patient's shoulder to raise the affected hand (Figure 2)

The guidelines of annotating performance components are described in Table I. The labels of *'ROM'* and *'Smoothness'* are annotated at the end of a motion on a three point scale. Those labels of *'ROM'* and *'Smoothness'* are converted to a binary label with the following mapping: a score 2 indicates a correct/normal performance component ($Y = 1$), and both score 1 and 0 describe 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) occurs or not.

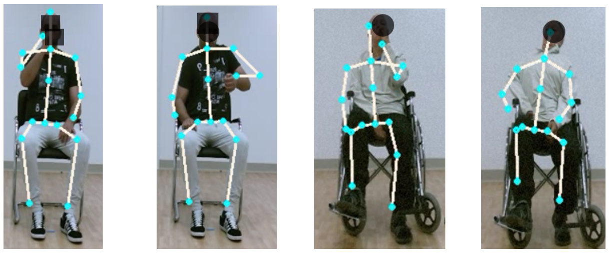

(a) Unaffected (b) Affected (c) Unaffected (d) Affected

Fig. 2: Two patients performing Exercise 1 with different compensated motions.

TABLE I: Guidelines to Assess Stroke Rehabilitation Exercises.

Performance Components	Labels	Guidelines
Range of Motion (ROM)	0 1 2	Does not or barely involve any movement Less than half way aligned with an *'Target'* position Movement achieves an *'Target'* position
Smoothness	0 1 2	Excessive tremor or not smooth coordination Movement influenced by tremor Smoothly coordinated movement
Compensation	0/1 0/1 0/1	Head in abnormal/normal alignment Spine in abnormal/normal alignment Shoulder in abnormal/normal alignment

IV. Interactive Approach of an Assistive Robot for Personalized Assessment and Feedback

This paper presents an interactive approach of a robotic exercise coaching system (Figure 1a), which combines machine learning (ML) and rule-based (RB) models to assess the performance of an exercise and tunes with patient's data to generate personalized feedback. A ML model of our approach aims to extract meaningful patterns from a large amount of data and support generic assessment on the patient's quality of motion. As such a generic ML model might not perform well on an unobserved new patient's motion with unique characteristics, our approach also integrates 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 recombined to complement a generic ML model with a weighted average, ensemble technique [9], [14] 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 present an ensemble voting method that accommodates predictions on multiple consecutive frames for more accurate frame-level assessment. In the following subsections, we describe the components of our approach: feature extraction, ML models, RB models, hybrid models, and an ensemble voting method.

A. Feature Extraction

This work represents an exercise motion with sequential joint coordinates from a Kinect v2 sensor (Microsoft, Redmond, USA) and extracts various kinematic features [19]. For the *'ROM'*, we compute joint angles (e.g. elbow flexion, shoulder flexion, elbow extension) and normalized relative trajectory (i.e. the Euclidean distance between two joints - head and wrist, head and elbow) [19]. For the *'Smoothness'*, we compute the following speed related features: speed and zero crossing ratio of acceleration [19]. As our work demonstrates the feasibility with upper-limb exercises, we computed these speed related features on wrist and elbow joints. For the *'Compensation'*, we compute normalized trajectories: distances between joint positions of head, spine, shoulder in x, y, z axis from the initial to current frame [19].

A moving average filter with the window size of five frames is applied to reduce noise of acquiring joint positions from a Kinect sensor similar to [19]. Given an exercise motion, we compute a feature matrix $\mathbf{F} = \{f_1, ..., f_T\} \in R^{T \times d}$ with T number of frames and d features and 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 summarized feature vector, $X \in R^{5d}$. This summarized feature vector is utilized for the assessment on *'ROM'* and *'Smoothness'* performance components and a feature matrix is utilized for the frame-level assessment on *'Compensation'* performance component.

B. Machine Learning (ML) Model

A machine learning (ML) model applies a supervised learning algorithm with training data from all patients except a patient for testing to predict the quality of motion or compute the score of being correct on a performance component (P_{ML}). We explore various supervised learning algorithms: a Decision Trees (DT), Linear Regression (LR), Support Vector Machine (SVM), a Neural Network (NN), and a Long Short Term Memory (LSTM) network using the *'Scikit-learn'* [20] and the *'PyTorch'* libraries [21].

For DTs, Classification and Regression Trees (CART) is utilized to build pruned trees. For LR models, we apply $L1, L2$ regularization or linear combination of $L1$ and $L2$ (ElasticNet with 0.5 ratio). For SVMs, we apply either linear, polynomial or Radial Basis Function (RBF) kernels with the penalty parameter, $C = 1.0$. NNs are trained while grid-searching over various architectures (i.e. one to three layers with $32, 64, 128, 256, 512$ hidden units) and different learning rates (i.e. $0.0001, 0.005, 0.001, 0.01, 0.1$). For LSTM networks, we have two architectures: (i) many-to-one (Figure 3a) for the *'ROM'* and *'Smoothness'* performance components, which leverages sequential kinematic features to assess performance components at the end of an exercise and (ii) many-to-many (Figure 3b) for the *'Compensation'* performance component that utilizes kinematic features at every frame of an exercise for frame-level assessment. We apply 0.5 drop-out to LSTM layers and explore various fully connected layers for LSTMs: one to three layers with $32, 64, 128, 256, 512$ hidden units) and different learning rates (i.e. $0.0001, 0.005, 0.001, 0.01, 0.1$). Fully connected layers of both NNs and LSTMs have applied *'ReLu'* activation functions. Both NNs and LSTMs are trained with cross-entropy loss and the mini-batch size of 1 and epoch 1.

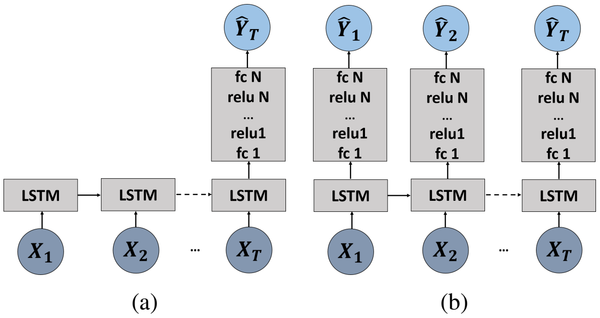


Fig. 3: Architectures of LSTM networks: a) many-to-one that utilizes sequential kinematic features to predict the quality of *'ROM'* and *'Smoothness'* at the end of a motion. (b) many-to-many that utilizes kinematic features at every frame of a motion to predict the quality of *'Compensation'*.

C. Rule-Based (RB) Model

A rule-based (RB) model leverages the set of feature-based rules from therapists to estimate the quality of a motion [9]. For an initial development of the RB model, semi-structured interviews are conducted with two therapists ($\mu = 5.0$, $\sigma = 1.05$ years of experience in stroke rehabilitation) to elicit their knowledge of assessing stroke rehabilitation exercises. The knowledge of therapists is formalized as 15 independent *if-then* rules. For example, the assessment on the ROM component for Exercise 1 is specified as follows [9]:

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

where $p(j, c)$ indicates a joint position 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, $p^{max}(wr, c_y)$, related to that of a spine shoulder joint, $p^{max}(spsh, c_y)$, 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}) can be computed with the following equation:

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

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. $p^{max}(spsh, c_y)$ 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 (Figure 1a). For the computation of personalized threshold values, we utilize held-out user's unaffected motions to learn a Gaussian probability density function $f(x_r) \sim N(\mu_r, \sigma_r^2)$, where x_r indicates the feature value of a rule r and μ_r and σ_r are 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. $\tau_r \in [\mu_r + 2\sigma_r, \mu_r + 3\sigma_r]$).

D. Hybrid Model

A hybrid model (HM) applies a weighted average, ensemble technique [8], [5] to combine machine learning (ML) and rule-based models to assess the quality of motion [9]. 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 $[0, 1]$). The equation of computing the score of being correct using the HM, P_{HM} is as follows:

$$P_{HM} = \frac{\rho_{ML}}{\rho_{ML} + \rho_{RB}} P_{ML} + \frac{\rho_{RB}}{\rho_{ML} + \rho_{RB}} P_{RB} \qquad (3)$$

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.

E. Ensemble Voting Method for Frame-Level Assessment

Our approach supports the detection of a compensation motion at frame-level so that a robotic exercise coaching system can provide a patient corrective feedback when an erroneous motion has occurred. However, such a frame-level assessment, identifying the exact boundaries of a compensation motion is challenging [22]. Thus, our approach applies an ensemble voting method that utilizes predictions on multiple consecutive V_f frames for more robust frame-level assessment. The process of this method is consist 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 t frame with an assessment model h (e.g. machine learning, rule-based or hybrid) and a feature vector f_t. The first process of an initial frame-level prediction runs in a continuous fashion 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:

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

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 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 $h(f_t)$. Through leveraging votes from past $V_f - 1$ frames to current t frame, our approach can support more robust frame-level assessment.

V. Experiments

A. 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) [19]. Fifteen post-stroke patients (2 females) with diverse functional abilities from mild to severe impairment (37 ± 21 out of 66 Fugl Meyer Scores [18]) 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 trajectory of joints and video frames at 30 Hz. The starting and ending frames of exercise movements were manually annotated.

Two therapists ($\mu = 5.0$, $\sigma = 1.0$ years of experience in stroke rehabilitation) annotated the dataset to implement our approach and compute expert's agreement level. They individually watched the recorded videos of patient's exercise movements and annotated the performance components of exercise motion dataset. For the frame level annotation of *'Compensation'* performance component, two expert annotators reviewed the images that are extracted from the recorded videos with the corresponding sampling frequency and the FFmpeg tool [23]. The annotations of experts are compared to measure expert's agreement on F1-scores (i.e. *'Expert'* in Table II and III). We utilize the annotation of an expert, who evaluated the functional abilities of patients with Fugl Meyer Assessment and had more experience as the ground truth.

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

- ***'Training Data'*** (Figure 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'*** (Figure 1a) includes 10 unaffected motions and 10 affected motions of a testing post-stroke subject.

B. Evaluation

We apply Leave-One-Subject-Out (LOSO) cross validation on post-stroke patients to evaluate our approach. A machine learning model (ML) is trained with data from all subjects except one testing post-stroke subject. An initial rule-based (RB) model is 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) are tested with affected motions of the left-out post-stroke patient. This process is repeated fifteen times to assess all post-stroke patients' affected motions. In addition, we analyze the effect of tuning a model with held-out unaffected motions of the left-out post-stroke subject. We also explore different numbers of multiple consecutive V_f frames on our ensemble voting method for frame-level assessment (i.e. $V_f = 1, ..., 30$). For the performance metric, this work utilizes a F1-score that computes the harmonic mean of precision and recall for a more realistic measure of a model.

VI. Results

Table II 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 explore various approaches: a decision tree (ML-DT), a linear regression model (ML - LR), a support vector machine (ML - SVM), a neural network (ML - NN), and a long short term memory network (ML-LSTM). The parameters of machine learning models (i.e. hidden layers/units and learning rates of neural networks and LSTM networks) that achieve the best F1-score during leave-one-subject-out (LOSO) cross validation are summarized in Table IV.

In addition, we present the performance of the initial rule-based model (RB-Init) from the interviews with therapists and that of the fine-tuned rule-based model (RB-tuned) after accommodating held-out user's unaffected motions to tune threshold values for personalized assessment. The parameters of rule-based models (i.e. the range of the threshold value 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 Exercise 1 and 2.

For hybrid models (HMs), we describe the performance of the initial hybrid model (HM-Init) that integrates the outperforming, machine learning model (i.e. Neural Networks - ML-NN) with the initial rule-based model (RB-Init) and that of the tuned hybrid model (HM-tuned) that combines the ML-NN with the 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) and outperform other algorithms ($p < 0.01$ using a paired t-test over three exercises and three performance components in Table III): Decision Trees (0.7014 average F1-score), Linear Regression (0.6785 average F1-score), Support Vector Machines (0.7055 average F1-score), and Long Short Term Memory Networks (0.6736 average F1-score). Although most machine learning practitioners consider sequential modeling with recurrent neural networks (RNNs), our results show that a neural network architecture (i.e. ML-NN) performs better to model the task of assessing both frame-level and overall quality of motion than recurrent neural networks (i.e. ML-LSTM) similar to prior work [24] that showed the absence of the advantage of RNNs on various tasks in practice (e.g. natural language and audio processing).

TABLE II: Performances (avg. ± std. of F1-scores) of machine learning (ML) models, rule-based (RB) models, hybrid models (HMs), and expert's agreement. ‡ indicates HM-Tuned performs statistically better than the compared method (pairwise t-tests at 99% significance level).

Algorithm	Exercise 1	Exercise 2	Exercise 3	Overall
ML-DT ‡	0.7308 ± 0.0584	0.6848 ± 0.2032	0.6887 ± 0.0805	0.7014 ± 0.0255
ML-LR ‡	0.7164 ± 0.0234	0.6323 ± 0.0877	0.6867 ± 0.0618	0.6785 ± 0.0426
ML-SVM ‡	0.7390 ± 0.0148	0.6441 ± 0.1053	0.7334 ± 0.0245	0.7055 ± 0.0533
ML-NN	0.8428 ± 0.0809	0.7549 ± 0.1026	0.7720 ± 0.0433	0.7899 ± 0.0466
ML-LSTM ‡	0.7509 ± 0.0346	0.5886 ± 0.0608	0.6813 ± 0.0574	0.6736 ± 0.0814
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**
Expert	0.7908 ± 0.2146	0.8222 ± 0.1534	0.7196 ± 0.1754	0.7775 ± 0.0526

The initial rule-based model (RB-Init) achieves the lowest performance: 0.5827 average F1-score over all exercises. According to the further analysis on the initial 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. 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 are annotated as normal and performed without the compensated shoulder joint, the shoulder joint of those motions is located around 20% of the initial positions and mis-classified as compensated motions. This indicates the importance of generating personalized rules for patients with various physical characteristics and functional abilities.

The initial hybrid model (HM-Init) achieves 0.7447 average F1-score over all exercises. As the initial rule-based model has the limited performance, the initial hybrid model (HM-Init) that integrates the ML model with Neural Networks (ML-NN) and the initial RB model (RB-Init) leads to slightly lower performance than that of the ML-NN (i.e. 0.7899 average F1-score over all exercises). However, the HM-Init still outperforms other ML models (e.g. ML-DT, ML-LR, ML-SVM, and ML-LSTM).

To evaluate the feasibility of tuning a model for personalized assessment, we update the threshold values of a rule-based model with held-out patient's unaffected motion (as described in Section IV-C) and implement the tuned rule-based model (RB-Tuned) and tuned hybrid model (HM-Tune) that integrates the ML-NN model with the RB-Tuned model. Both RB-Tuned and HM-Tuned models significantly improve their performance to replicate therapist's assessment ($p < 0.01$ using the paired t-tests over three performance components of three exercises in Table III). Specifically, the RB model significantly improves 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)

TABLE III: Performances (average ± standard deviation of F1-scores) of various methods (i.e. machine learning, rule-based, and hybrid models) to assess the quality of motion. Best results of each column are boldfaced. ‡ indicates that the HM-Tuned performs statistically better than a compared method (pairwise t-tests at 99% significance level).

Algorithm		Exercise 1			Exercise 2			Exercise 3	
	ROM	Smooth	Comp	ROM	Smooth	Comp	ROM	Smooth	Comp
ML-DT ‡	0.7192 ± 0.3968	0.6791 ± 0.3865	0.7942 ± 0.0661	0.8919 ± 0.2448	0.4857 ± 0.4178	0.6767 ± 0.1116	0.6037 ± 0.3963	0.6986 ± 0.3499	0.7638 ± 0.0682
ML-LR ‡	0.6912 ± 0.4247	0.7375 ± 0.4048	0.7205 ± 0.1136	0.7335 ± 0.3732	0.5787 ± 0.4789	0.5848 ± 0.2122	0.7139 ± 0.3876	0.7302 ± 0.3771	0.6160 ± 0.0754
ML-SVM ‡	0.7251 ± 0.4136	0.7375 ± 0.4048	0.7545 ± 0.0776	0.7655 ± 0.3475	0.5787 ± 0.4789	0.5880 ± 0.2078	0.7593 ± 0.3811	0.7302 ± 0.3771	0.7107 ± 0.0642
ML-NN	0.9324 ± 0.1350	**0.7751** ± **0.3600**	0.8210 ± 0.0755	0.8696 ± 0.2852	0.6721 ± 0.4642	0.7229 ± 0.1354	0.7220 ± 0.2853	**0.7969** ± **0.3272**	**0.7970** ± **0.0345**
ML-LSTM ‡	0.7902 ± 0.3968	0.7375 ± 0.4048	0.7249 ± 0.1073	0.5333 ± 0.4989	0.5787 ± 0.4789	0.6537 ± 0.1425	0.6568 ± 0.4659	0.7469 ± 0.3645	0.6403 ± 0.0577
RB-Init ‡	0.8432 ±0.3094	0.4344 ± 0.3910	0.5669 ± 0.4340	0.8466 ± 0.2886	0.4950 ± 0.4094	0.6705 ± 0.4173	0.5320 ± 0.4632	0.2265 ± 0.3140	0.6294 ± 0.3590
RB-Tuned	0.9091 ± 0.2471	0.7524 ± 0.3494	0.8337 ± 0.0336	**0.9405** ± **0.1382**	**0.7044** ± **0.3704**	0.7579 ± 0.0365	0.7797 ± 0.3388	0.7302 ± 0.3771	0.7531 ± 0.0511
HM-Init ‡	0.9148 ± 0.1636	0.7680 ± 0.3648	0.7379 ± 0.3867	0.8419 ± 0.3156	0.5787 ± 0.4789	0.6975 ± 0.4269	0.8035 ± 0.2335	0.7267 ± 0.3748	0.6335 ± 0.4500
HM-Tuned	**0.9714** ± **0.0555**	0.7680 ± 0.3648	**0.8410** ± **0.0821**	0.9050 ± 0.2215	0.6419 ± 0.4605	**0.7837** ± **0.0647**	**0.9616** ± **0.0700**	0.7424 ± 0.3558	0.7963 ± 0.0221
Expert	0.9587 ± 0.0489	0.5490 ± 0.0011	0.8646 ± 0.2127	0.9630 ± 0.0427	0.6588 ± 0.1384	0.8449 ± 0.1408	0.7342 ± 0.2418	0.5373 ± 0.1148	0.8872 ± 0.1021

also significantly improves its performance around 11% from 0.7447 to 0.8235 average F1-scores over all exercises ($p < 0.01$) and outperform other approaches. The performance of the tuned hybrid model (HM-tuned) is better than those of the machine learning model with Neural Networks (ML-NN) and the RB-tuned around 0.03 average F1-score (i.e. 4% and 3% improvement respectively without statistical significance).

To analyze the effect of our ensemble voting method for frame-level assessment, we utilize the ML-NN, RB-Tuned, and HM-Tuned models and plot their average performance of detecting frame-level compensation on head, spine, shoulder joints over three exercises with various numbers of consecutive frames (V_f = 1, ..., 30). In Figure 4, all three models (i.e. ML-NN, RB-Tuned, HM-Tuned) improve their performance while leveraging prediction from multiple frames and achieve their best performance with $V_f = 29$. When we compare the performance of a model without and with an ensemble voting method ($V_f = 1$ and $V_f = 29$), the ML-NN model improves 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 improves its performance from 0.7655 ($V_f = 1$) to 0.7816 ($V_f = 29$) average F1-score ($p < 0.01$); the HM-Tuned model improves its performance from 0.7975 ($V_f = 1$) to 0.8070 ($V_f = 29$) average F1-score ($p < 0.01$).

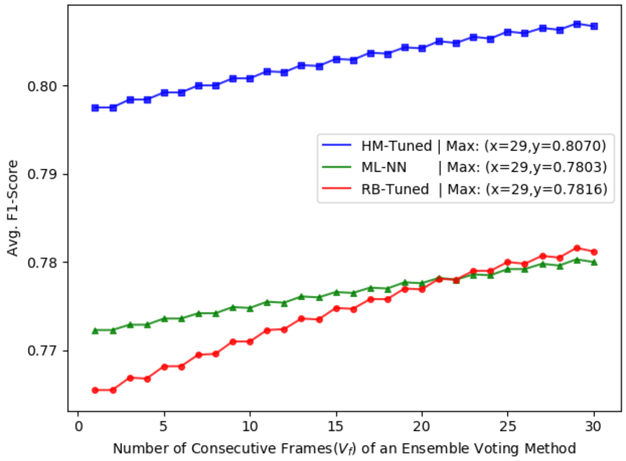


Fig. 4: Performance of frame-level assessment with different values of consecutive frames (V_f) using the tuned rule-based model (RB-Tuned), machine learning model with neural networks (ML-NN), tuned hybrid model (HM-Tuned)

VII. Discussion

Among various approaches, the machine learning model with Neural Networks (ML-NN), tuned rule-based model (RB-tuned), and the initial and tuned hybrid models (HM-Init and HM-tuned) 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, a RB model could be considered as a natural starting point to develop a robotic exercise coaching system that can assess the quality of motion and generate corrective feedback on patient's exercises [4], [5]. However, a RB model with generic threshold values (e.g. RB-Init and [6], [7], [9], [11], [13]) 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 a RB model with individual's held-out unaffected motions to derive personalized threshold values for assessment and corrective feedback.

As a robotic 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 of assessing exercises from data. However, it is not recommended to simply replace a rule-based (RB) model with a ML model using a complex, black-box algorithm. For instance, given a patient's affected motion that is incorrectly performed with compensation, a ML model with Neural Networks can just notify whether the compensation has occurred or not. 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 Figure 2b. Such feature-level analysis can be realized into the following personalized corrective feedback: *"Keep your head straight and do not raise your shoulder"*. 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.

VIII. Conclusion and Future Work

In this paper, we present an interactive approach with an ensemble voting method for a robotic exercise coaching system that integrates a data-driven machine learning model with an interpretable rule-based model and tunes with patient's data for transparent, personalized interaction with corrective feedback on patient's stroke rehabilitation exercises. This work shows that interactive approach achieves good agreement with expert's annotation, and discusses the importance of an interactive approach to tune a model with patient's motions for personalized assessment and an ensemble voting method to improve frame-level assessment.

For automated assessment capability of a robotic exercise coaching system, an interactive rule-based model can be regarded as a starting point. In addition, after data collection, our work emphasizes the importance of creating a model with interpretability instead of just applying a complex deep learning model. We discuss that a hybrid model can augment a rule-based model with new insights on data from a machine learning model, but also support transparent and personalized interaction of a robotic exercise coaching system.

For the interaction with a patient, we utilize our presented approach and implement a robotic coaching system that can show the tracked joints of a patient's exercise motion and predicted assessment in a visualization interface, but generate real-time audio corrective feedback and gestures from a robot (Figure 1c). In future, we plan to explore the effect of personalized corrective feedback from a robotic exercise coaching system to support therapeutic rehabilitation sessions with post-stroke patients.

APPENDIX

TABLE IV: Parameters of Machine Learning Models

	Hidden Layers and Units / Learning Rate		
	ROM	Smoothness	Comp
E1	- NN: (256, 256, 256) / 0.005 - LSTM: (16, 16, 16) / 0.0001	- NN: (16) / 0.0001 - LSTM: (16) / 0.005	- NN: (512, 512, 512) / 0.005 - LSTM: (16) / 0.005
E2	- NN: (32, 32, 32) / 0.01 - LSTM: (16) / 0.0001	- NN: (32) / 0.0001 - LSTM: (16) / 0.0001	- NN: (256, 256) / 0.0001 - LSTM: (16) / 0.0001
E3	- NN: (16) / 0.005 - LSTM: (32) / 0.0001	- NN: (128) / 0.0001 - LSTM: (16) / 0.005	- NN: (256, 256, 256) / 0.1 - LSTM: (16, 16, 16) / 0.0001

ACKNOWLEDGMENT

The authors thank all participants in the study. 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)] and the National Science Foundation (NSF) under grant number CNS-1518865.

References

[1] S. B. O'Sullivan, T. J. Schmitz, and G. Fulk, *Physical rehabilitation*. FA Davis, 2019.

[2] J. H. Rimmer, E. Wang, and D. Smith, "Barriers associated with exercise and community access for individuals with stroke." *Journal of rehabilitation research & development*, vol. 45, no. 2, 2008.

[3] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.

[4] M. J. Matarić, J. Eriksson, D. J. Feil-Seifer, and C. J. Winstein, "Socially assistive robotics for post-stroke rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 4, no. 1, p. 5, 2007.

[5] M. H. Lee, "Intelligent agent for assessing and guiding rehabilitation exercises," in *IJCAI*, 2019, pp. 6444–6445.

[6] J. Fasola and M. J. Matarić, "A socially assistive robot exercise coach for the elderly," *Journal of Human-Robot Interaction*, vol. 2, no. 2, pp. 3–32, 2013.

[7] B. Görer, A. A. Salah, and H. L. Akın, "An autonomous robotic exercise tutor for elderly people," *Autonomous Robots*, vol. 41, no. 3, pp. 657–678, 2017.

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[9] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. Bermúdez i Badia, "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*, ser. UMAP '20. ACM, 2020, p. 303–307.

[10] D. Feil-Seifer and M. J. Mataric, "Defining socially assistive robotics," in *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, 2005, pp. 465–468.

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

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

[13] P. Tanguy, O. Rémy-Néris, *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)*. IEEE, 2016, pp. 1138–1143.

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

[15] M. H. Lee, "An intelligent decision support system for stroke rehabilitation assessment," in *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2019, pp. 694–696.

[16] V. L. Feigin, B. Norrving, and G. A. Mensah, "Global burden of stroke," *Circulation research*, vol. 120, no. 3, pp. 439–448, 2017.

[17] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, *et al.*, "Opportunities of a machine learning-based decision support system for stroke rehabilitation assessment," *arXiv:2002.12261*, 2020.

[18] J. Sanford, J. Moreland, L. R. Swanson, P. W. Stratford, and C. Gowland, "Reliability of the fugl-meyer assessment for testing motor performance in patients following stroke," *Physical therapy*, vol. 73, no. 7, pp. 447–454, 1993.

[19] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. B. i. Badia, "Learning to assess the quality of stroke rehabilitation exercises," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 218–228.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[22] M. Hasan and A. K. Roy-Chowdhury, "Continuous learning of human activity models using deep nets," in *European conference on computer vision*. Springer, 2014, pp. 705–720.

[23] F. Developers, "ffmpeg tool," *URL: http://ffmpeg. org*, 2016.

[24] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv:1803.01271*, 2018.