Combining Human and Machine Intelligence to Assess Stroke Rehabilitation Exercises*

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

Abstract

Advanced machine learning algorithms and sensors provide the opportunity of automatically assessing rehabilitation exercises to assist therapist's management on patient's rehabilitation programs. However, despite recent advances in machine learning, AI systems cannot perfectly reproduce therapist's assessment on patients' diverse conditions. This paper presents an interactive hybrid approach that can complement an imperfect machine learning model with a rule-based model for more accurate rehabilitation assessment, and demonstrates the effect of accommodating therapist's feedback on an AI-based system to assess stroke rehabilitation exercises. This work discusses the importance of augmenting an imperfect machine learning model with expert's knowledge for more accurate decision making.

1. Introduction

Patients with musculoskeletal and neurological disorders (e.g. stroke) require physical rehabilitation programs for several months to restore their functional ability and enhance their quality of life. During a rehabilitation program, therapists assess patient's functional status and provide corrective feedback. As therapists cannot observe all exercise trials of a patient, they prescribe home exercise regimens (O'Sullivan et al., 2019). In the follow-up visits, therapists rely on a patient's self-report to discuss patient's progress and decide how to adjust exercise regimens (O'Sullivan et al., 2019). However, therapists have difficulty with making informed decision on adjusting treatment interventions without observing patient's exercises or quantitative exercise data (Jones et al., 2006).

Advanced machine learning and sensor technologies have the potential of computer-assisted rehabilitation systems that automatically monitors and assesses patient's status to support patient's in-home physical rehabilitation (Webster & Celik, 2014). Previous work on computer-assisted rehabilitation monitoring and assessment can be categorized into rule-based and machine learning approaches (Siewiorek et al., 2012; Lee et al., 2020b). A rule-based approach derives a set of monitoring rules through the involvement of experts in the design process (Lee et al., 2020b). However, it is difficult to properly articulate an expert's decision making process on a complex monitoring task. Alternatively, a machine learning model with labeled sensor data can automatically extract a meaningful function (e.g. Neural Network model) to assess the quality of motion (Patel et al., 2010; Das et al., 2011; Lee et al., 2019). However, it is challenging to derive a model that can reproduce therapist's assessment for patients with different physical characteristics. In addition, when a model with complex algorithms fails to correctly assess rehabilitation exercises and does not provide any explanations to support therapist's decision making, therapists can lose trust and abandon it (Kizilcec, 2016; Khairat et al., 2018).

In this paper, we describe and evaluate an interactive hybrid approach (Lee et al., 2020a) that integrates a machine learning model with a rule-based model from therapists to assess the quality of motion (Figure 1). This work utilizes the dataset of three upper-limb rehabilitation exercises from 15 post-stroke and 11 healthy subjects with the corresponding assessment scores by expert therapists (Lee et al., 2019). From this dataset, we apply reinforcement learning to identify the most important features for assessment and learn a machine learning (ML) model to predict the scores on the exercises of patients using leave-one-subject-out cross validation. For the development of an initial rule-based (RB) model, we conducted a semi-structured interview with therapists to elicit their knowledge of assessing rehabilitation exercises. A ML model and a RB model are integrated with a weighted average ensemble technique (Baltrušaitis et al., 2019) to derive the Hybrid Model (HM) (Lee et al., 2020b) for assessment.

[1] Carnegie Mellon University, Pittsburgh, PA, USA [2] Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal [3] Madeira Interactive Technology Institute, University of Madeira, NOVA-LINCS, Funchal, Portugal. Correspondence to: Min Hun Lee <minhunl@cs.cmu.edu>. *A detailed version of this paper is published at the ACM Conference on Health, Inference, and Learning (CHIL) 2020.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

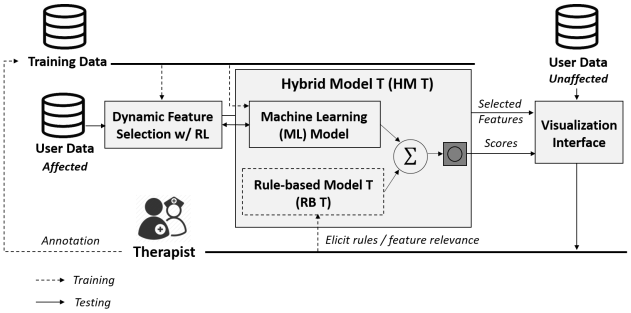


Figure 1. Flow diagram of an interactive hybrid approach for rehabilitation assessment that combines machine learning (ML) models with a rule-based (RB) model. ML models automatically select features and predict the quality of motion to generate a patient-specific analysis. A therapist can review this analysis and provide feedback on feature relevance to tune a model for personalized rehabilitation assessment

Once a new patient performs the exercise with patient's unaffected and affected side, the visualization interface of our approach (Figure 2) shows a patient-specific analysis: the predicted quality of motion on three performance components (i.e. *'Range of Motion'*, *'Smoothness'*, *'Compensation'*) and the comparison between unaffected and affected sides on the most important kinematic features (Lee et al., 2020c). Therapists can review this analysis to better understand patient's performance and provide feedback on feature relevance to tune a model for personalized rehabilitation assessment.

After implementing our approach, we performed a user study with therapists to evaluate our approach and explore the effect of accommodating therapist's feedback for personalized rehabilitation assessment. Our experimental results demonstrate that our approach can iteratively elicit and accomodate therapist's feature-based feedback to significantly improve its performance of replicating therapist's assessment from 0.8377 to 0.9116 average F1-scores on three exercises ($p < 0.01$) (Lee et al., 2020a).

Although prior work demonstrated the feasibility of monitoring and assessing rehabilitation exercises (Webster & Celik, 2014), there is a lack of evaluations on how such imperfect technologies can be utilized by therapists. This paper contributes to increase knowledge on an interactive machine learning approach that can accommodate therapist's feedback for more accurate rehabilitation assessment.

2. Stroke Rehabilitation as a Test Domain

Stroke is the second leading cause of death and third most common contributor to disability (Feigin et al., 2017). Thus, we selected stroke rehabilitation as a probe domain. We had iterative discussion with three therapists ($\mu = 6.33$, $\sigma = 2.05$ years of experience in stroke rehabilitation) to specify the designs of our study on stroke rehabilitation: exercises and performance components for assessment (Lee et al., 2020c).

2.1. Three Task-Oriented Upper Limb Exercises

This paper utilizes three upper-limb stroke rehabilitation exercises recommended by therapists (Lee et al., 2020c). For Exercise 1, a subject has to raise subject's wrist to the mouth as if drinking water. For Exercise 2, a subject has to pretend touching a light switch on the wall. For Exercise 3, a subject has to practice the usage of a cane by extending subject's elbow in the seated position.

2.2. Performance Components

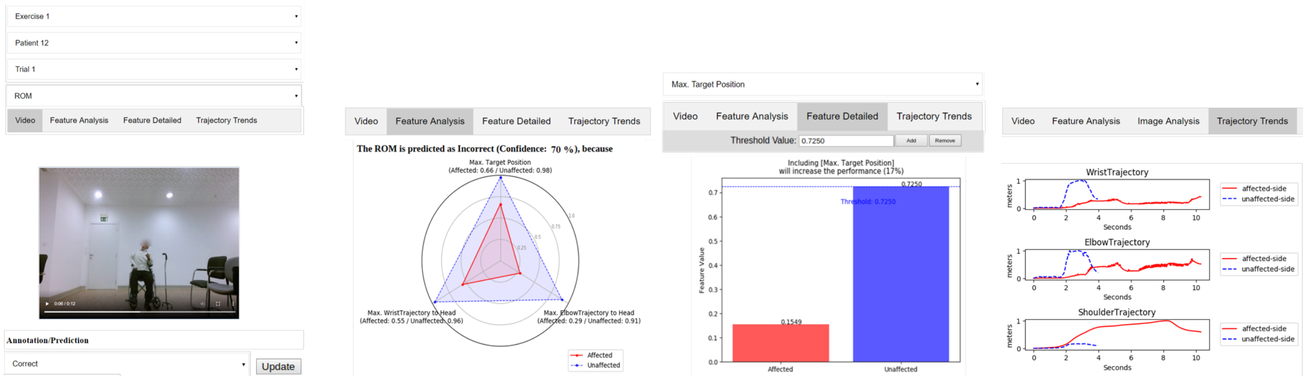
After reviewing commonly used stroke assessment tools (i.e. the Fugl Meyer Assessment (Sanford et al., 1993) and the Wolf Motor Function Test (Taub et al., 2011)) and having iterative discussion with therapists, we specified three common performance components to assess the quality of motion: *'Range of Motion (ROM)'*, *'Smoothness'*, and *'Compensation'*. For binary labels, we refer a correct/normal performance component to $Y = 1$, and an incorrect/abnormal performance component to $Y = 0$.

The *'ROM'* component describes how closely a patient performs a task-oriented exercise. The *'Smoothness'* component indicates the degree of trembling and irregular movement of joints while performing an exercise. The *'Compensation'* component checks whether a patient performs any compensated movements to achieve a target movement. For instance, a patient might elevate his/her shoulder to raise the affected hand (Lee et al., 2020c).

2.3. Kinematic Features

This work represents an exercise motion with sequential joint coordinates from a Kinect v2 sensor (Microsoft, Redmond, USA) and extracts various kinematic features (Lee et al., 2019).

For the *'ROM'* component, we compute joint angles (e.g. elbow flexion, shoulder flexion, elbow extension), and normalized relative trajectory (i.e. Euclidean distance between two joints - head and wrist, head and elbow). For the *'Smoothness'* component, we compute various speed related features: the speed, acceleration, jerk, zero crossing ratio of acceleration and jerk, and Mean Arrest Period Ratio (the portion of the frames when speed exceeds 10% of the maximum speed) (Lee et al., 2019). As we have upper-limb exercises, we computed these speed related features on wrist and elbow joints. For the *'Compensation'* component, we compute joint angles (i.e. the elevated angle of a shoulder, the tilted angle of


Exercise 1

Patient 12

Trial 1

ROM

Annotation/Prediction

Correct

The ROM is predicted as Incorrect (Confidence: 70 %), because

Max. Target Position

Threshold Value: 0.7250

Including [Max. Target Position] will increase the performance (17%)

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

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

head, spine, and shoulder abduction) and normalized trajectories (the distance between joint positions of head, spine, shoulder joints in x, y, z axis from the initial to the current frames) to distinguish a compensated movement.

Before extracting features, we apply a moving average filter with the window size of five frames to reduce noise of acquiring joint positions from a Kinect sensor similar to (Lee et al., 2019). For each exercise motion, we compute a feature matrix ($\mathbf{F} \in R^{t \times d}$) with t frame and d features, and statistics (i.e. max, min, range, average, and standard deviation) over all frames of the exercise to summarize a motion into a feature vector ($X \in R^{5d}$).

3. Interactive Hybrid Approach

Machine learning approaches make great progress in various fields that can afford a large dataset (e.g. speech recognition (Amodei et al., 2016) and autonomous vehicles (Xu et al., 2017)). However, the healthcare domain often involves a small dataset, which makes the application of automated approaches difficult or even impossible (Holzinger, 2016). Interactive machine learning approach seems to be a promising approach while making use of human cognitive abilities (Fails & Olsen Jr, 2003; Kulesza et al., 2015).

This paper presents an interactive hybrid approach (Figure 1), which aims at integrating the benefits of a machine learning (ML) model and a rule-based (RB) model from therapist's feedback (Lee et al., 2020a). This approach can automatically identify the most salient features to predict the quality of motion and generate a patient-specific analysis that compares those identified features of patient's affected and unaffected motions. This patient-specific analysis can assist therapists to gain new insights on patient's exercise motions and elicit their domain knowledge on feature relevance. Utilizing elicited feature relevance, our approach can iteratively update a rule-based (RB) model for personalized rehabilitation assessment. In the following subsections, we describe the components of our approach: dynamic feature selection using reinforcement learning, a ML model, a RB model, Hybrid Model (HM), and visualization interface.

3.1. Feature Selection

Kinematic feature analysis is an important source for therapists to quantitatively and objectively understand patient's performance (Wu et al., 2000). Yet, simply presenting all features can overwhelm therapists and limit therapist's ability to gain insights on patient's performance. As therapists have limited availability to administrate multiple patients, therapists should minimize the amount of time on analyzing kinematic features while accurately diagnosing patient's status. Thus, we aim at automatically identifying salient features of assessment to learn a sparse machine learning (ML) model and generate an interpretable and succinct patient-specific report.

The classical approaches of feature selection (e.g. filter, wrapper, embedded methods) (Tang et al., 2014) find a fixed feature set to the entire training dataset for all patients. In contrast, this paper applies a Markov Decision Process (MDP) (Janisch et al., 2019) to find the optimal feature set for each patient's motions. As each patient has different physical and functional status, we hypothesize and demonstrate that feature selection with MDP can perform better than a classical feature selection approach for personalized rehabilitation assessment (Lee et al., 2020a).

3.1.1. Problem Definition

We formulate this problem of feature selection as Markov Decision Process (MDP), where each episode is to classify an instance and the environment is the power set of the feature space. An agent sequentially determines whether to query an additional feature or classify a sample while receiving a negative reward for recruiting a feature or mis-classification. To solve this problem, we apply Deep Q-network with Double Q-learning (Mnih et al., 2015; Van Hasselt et al., 2016).

We mathematically describe the Markov Decision Process (MDP) with similar notations of (Dulac-Arnold et al., 2011; Janisch et al., 2019) as follows:

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

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

3.1.2. Implementation Details

We utilize *'PyTorch'* libraries (Paszke et al., 2017) to implement a neural network with parameters θ (Q_θ) for deep Q-learning (Mnih et al., 2015). The input layer of the network consists of feature and binary mask vectors (Janisch et al., 2019). This masking input vector is to indicate whether a feature is recruited or not. Specifically, we let $m \in \{0, 1\}^n$ be an n-dimensional vector for an environment of n features, where $m_i = 1$ if the agent has queried feature i thus far in the episode and 0 otherwise. The target network is also used for a machine learning (ML) model. The architectures and parameters of the Neural Networks are described in the Table 2.

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

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

where $r(s, a, s')$ indicates the received reward and γ indicates the discounted factor. We clip a gradient if a gradient norm exceeds 1.0 (Janisch et al., 2019) and update the target network after each step. Instead of directly updating the weight of the target network, we apply soft target updates (Lillicrap et al., 2015): $\theta' \leftarrow \rho\theta + (1 - \rho)\theta'$, where $\theta \leq 1$. ρ denotes this soft target update factor and is specified as 0.1. This soft target updates can improve the stability of learning parameters of target networks. As the application of soft target updates may lead to slow learning, we apply an experience replay (Mnih et al., 2015) for sampling efficiency. Specifically, the environment with randomly drawn samples is simulated and the transition data is recorded to the experience replay buffer. As the environment is episodic with a short length, we choose a value 1.0 for the discount factor γ. In addition, we apply the ϵ-greedy policy to control the exploration. Specifically, we linearly decrease the ϵ value from the $\epsilon_{start}(0.5)$ to the $\epsilon_{end}(0.05)$ with a step value, $\epsilon_{step}(0.02)$.

3.2. Machine Learning (ML) Model

A machine learning (ML) model utilizes a supervised learning algorithm and training data from all patients except a patient for testing to predict the quality of motion or compute a score of being correct on a performance component, $P_{ML} = P(Y = 1|X)$. We utilize a Neural Network (NN) due to its outpeformance as shown in (Lee et al., 2019). The parameters of NNs are grid-searched over various architectures (i.e. one to three layers with $32, 64, 128, 256, 512$ hidden units) and different initial learning rates (i.e. $0.0001, 0.005, 0.001, 0.01, 0.1$) to have highest leave-one-subject-out cross-validation performance. NN models applies the *'ReLu'* activation functions and *'AdamOptimizer'* and are trained until the tolerance of optimization is 0.0001 or the maximum 200 iterations.

3.3. Rule-based (RB) Model

A rule-based (RB) model utilizes the set of feature-based rules from therapists to estimate the quality of a motion (Lee et al., 2020b). For the initial development of a rule-based model, we conducted a semi-structured interview with two therapists to elicit their knowledge of assessing stroke rehabilitation exercises. The expert knowledge of therapists is formalized as 15 independent *if-then* rules. For example, the rule of the *'ROM'* for Exercise 1 is specified as follows:

$$\hat{Y} = \begin{cases} 1 & \text{if } p^{max}(wr, c_y) >= p^{max}(spsh, c_y) \\ 0 & \text{else} \end{cases} \quad (2)$$

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 compares the maximum position of wrist joint, $p^{max}(wr, c_y)$ with that of 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.

The score of being correct on each performance component using a rule-based (RB) model can be computed using the following equation (Lee et al., 2020a):

$$P_{RB^T} = \frac{1}{|\mathbb{R}^T|} \sum_{r \in R^T} \min(\frac{f_r}{\tau_r}, 1) \quad (3)$$

where f_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. $p^{max}(sh, c_y)$ for the example above). $\mathbb{R}^T$ describes the set of rules considered relevant by the therapists with T number of interactions. $\min$ 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.

Furthermore, a rule-based (RB) model can be iteratively updated with the elicited expert's feedback on salient features. Given the identified salient features, our approach can generate a user-specific feature analysis between patient's unaffected and affected movements as shown in Figure 2. For the example of Figure 2, a user-specfic analysis describes that *'Incorrect'* ROM is predicted due to smaller maximum target position (affected: 0.66, unaffected: 0.98), maximum wrist trajectory to head (affected: 0.55 / unaffected: 0.96), and maximum elbow trajectory to head (affected: 0.29 / unaffected: 0.91). After reviewing a user-specific feature analysis, therapists can indicate whether presented features should be included or excluded (Kulesza et al., 2015) for assessment. When including a feature, therapists have an option of either utilizing the feature value of the unaffected side or specifying a value for the threshold value (τ_r) of a feature-based rule.

3.4. Hybrid Model

A hybrid model (HM) applies a weighted average, ensemble technique (Baltrušaitis et al., 2019) to integrate two perspectives on assessment: a machine learning (ML) model and a rule-based (RB) model from therapists (Lee et al., 2020b).

For the classification of the quality of motion, a hybrid model (HM) computes the weighted average of prediction scores from two models, in which the contribution of each model is weighted by the performance of a model (i.e. the F1-score of each model in the range of $[0, 1]$). Specifically, we compute the prediction score of HM, P_{HM^T} as follows:

$$P_{HM^T} = \frac{\rho_{ml}}{\rho_{ml} + \rho_{rb^T}} P_{ML} + \frac{\rho_{rb^T}}{\rho_{ml} + \rho_{rb^T}} P_{RB^T} \quad (4)$$

where T indicates the number of interaction/feedback from therapists, P_{ML} and P_{RB^T} indicate the scores of a machine learning (ML) model and a rule-based (RB) model at the T iteration respectively, and ρ_{ml} and ρ_{rb^T} describe the F1-scores of ML and RB models at T-th interactions.

3.5. Visualization Interface

Based on the prior work that describes the needs of therapists during stroke rehabilitation assessment (Lee et al., 2020c) and the guidelines of Human Artificial Intelligence (AI) interaction (Kulesza et al., 2015; Amershi et al., 2019), we implemented the web-based visualization that presents the predicted quality of performance components (e.g. *'Range of Motion', 'Smoothness', 'Compensation'*) as well as an explanation on the prediction of a model, a user-specific analysis that contains feature analysis, detailed feature values, and trajectory trends (Figure 2).

According to the focus-group discussions with therapists, they desire quantitative feature analysis for more accurate assessment instead of repetitively watching a video of patient's exercise motions and solely relying on their own knowledge and experience (Lee et al., 2020c). To present *"contextually relevant information"* (Amershi et al., 2019) for the assessment, this interface presents a video of patient's exercise motion along with a patient-specific analysis that includes predicted quality of motion, feature analysis (Figure 2b and 2c), and trajectory trends (Figure 2d).

To *"make clear how well the system can do"* (Amershi et al., 2019), the performance of a system is also included when presenting the predicted quality of performance components (Figure 2b and 2c).

As therapists utilize patient's unaffected motion as normality to assess patient's performance (Lee et al., 2020c), this interface follows this current practice, *"social norms"* (Amershi et al., 2019), and includes the comparison between the affected and unaffected side to present salient features (Figure

2b) and trajectory trends of three major joints (e.g. shoulder, elbow, and wrist) for upper-limb exercises (Figure 2d).

To *"avoid overwhelming"* (Kulesza et al., 2015) therapists, this interface presents only three salient features for each performance component with the highest information gain. A radar chart is utilized to effectively present multivariate data (Saary, 2008).

In addition, our interface supports to *"honor user feedback"* (e.g. feature-based feedback) (Kulesza et al., 2015). A feature-based feedback indicates the relevance of an identified feature for the assessment or the specification of a threshold value to generate a feature-based rule for personalized rehabilitation assessment. We present the changes in the performance of a model to support therapist's decision making (e.g. *"Including Max. Target Position will increase the performance (17%)"* in Figure 2c).

4. Experiments

4.1. Dataset of Three Upper-Limb Exercises

The dataset of three exercises is collected from 15 post-stroke and 11 healthy subjects using a Kinect v2 sensor (Lee et al., 2019). During the data collection, a sensor was located at the 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.

Fifteen post-stroke patients (13 males and 2 females) participated in two sessions for data collection. During the first session, a therapist evaluated post-stroke patient's functional ability using a clinically validated tool, the Fugl Meyer Assessment (FMA) (maximum score 66 points) (Sanford et al., 1993). 15 stroke survivors had diverse functional abilities from mild to severe impairment (37 ± 21 Fugl Meyer Scores). During the second session, a stroke patient performed 10 repetitions of each exercise with both affected and unaffected sides. Eleven healthy subjects (10 males and 1 female) performed 15 repetitions with their dominant arms for each exercise.

We divide the collected data into *'Training'* and *'User'* data. ***'Training Data'*** (Figure 1) is composed of 165 unaffected motions from 11 healthy and 140 affected motions from 14 post-stroke subjects to train a feature selection model and a machine learning (ML) model.
'User Data' (Figure 1) includes each stroke survivor's unaffected and affected motions. Both unaffected and affected motions of a testing post-stroke subject are excluded to train machine learning (ML) models. Using testing subject's affected motions, our approach dynamically selects subject-specific features and predicts the quality of motion on performance components. Both unaffected and affected motions of a testing subject are utilized to generate a patient-specific analysis of the visualization interface (Figure 2).

Two therapists annotated the dataset to implement our approach and compute the baseline agreement level of therapists. They individually watched the recorded videos of patient's exercise movements (Figure 2a) and annotated the performance components of exercise motion dataset without analysis of our system (Figure 2d, 2c, 2d). For implementation, we utilize the annotation of a therapist, who evaluated the functional ability of patients with Fugl Meyer Assessment, as ground truth. The annotation of the other therapist is compared with the ground truth annotation to compute therapist's agreement on F1-scores (TPA in Table 1).

4.2. Study with Therapists

To evaluate the feasibility of our interactive hybrid approach, we recruited five therapists with $\mu = 4.00$, $\sigma = 1.67$ years of experience in stroke rehabilitation and analyzed the effect of therapist's feedback on the system performance to predict rehabilitation assessment.

After signing the IRB approved consent form, each therapist was instructed on the task of providing feature-based feedback with dummy data. Specifically, feature-based feedback includes the following three options: 1) include or 2) remove a selected feature for assessment, or 3) updating the threshold value of a selected feature for assessment. For the task, each therapist was asked to provide feature-based feedback to make the predicted quality of motion from the interface as accurate as possible during a 30 minutes session. We assigned non-overlapping, three patients for each therapist to generate feature-based feedback on all post-stroke survivors in our dataset.

5. Results

5.1. Implementation

We apply Leave-One-Subject-Out (LOSO) cross validation on post-stroke patients to evaluate the implementation of our approach. A model is trained with data from all subjects except one post-stroke survivors and is tested with affected motions of the left-out post-stroke survivor. This process is repeated fifteen times to evaluate all post-stroke subjects' affected motions. To generate patient-specific analysis, held-out unaffected and affected motions of the left-out post-stroke survivor are utilized. For the performance metric, we utilize a F1-score.

Table 1 summarizes the performance of our approach, which measures the agreement with therapist's assessment using average F1-scores on the three exercises. The parameters of NNs (i.e. hidden layers/units and learning rate) that achieve the best F1-score during leave-one-subject-out cross

Table 1. Performance (F1-scores) of machine learning (ML) models, rule-based (RB) models, and hybrid models (HMs), and therapist's agreement (TPA). *** indicates that HM 10 performs significantly better than the compared method using paired t-tests at 99% significance level.

	Exercise 1 (E1)	Exercise 2 (E2)	Exercise 3 (E3)	Overall
ML - RL***	0.8331± 0.0059	0.7973 ± 0.0867	0.8053 ± 0.0496	0.8119 ± 0.0526
ML - RFE***	0.6742 ± 0.0715	0.7628 ± 0.1708	0.6415 ± 0.0806	0.6928 ± 0.1147
ML - NN***	0.8632 ± 0.0816	0.8388 ± 0.0518	0.7818 ± 0.0096	0.8279 ± 0.0605
RB 1***	0.6148 ± 0.1702	0.6932 ± 0.1630	0.4384 ± 0.1569	0.5821 ± 0.1066
RB 10***	0.7787 ± 0.1315	0.7607 ± 0.0872	0.7533 ± 0.0079	0.7642 ± 0.0106
HM 1***	0.8684 ± 0.0576	0.8159 ± 0.1195	0.8073 ± 0.0620	0.8305 ± 0.0270
HM 10	**0.9329 ± 0.0266**	**0.9218 ± 0.0539**	**0.8802 ± 0.0453**	**0.9116 ± 0.0226**
TPA***	0.8120 ± 0.1458	0.7790 ± 0.1324	0.7654 ± 0.1382	0.7854 ± 0.0195

validation are summarized in the Table 2.

Machine learning (ML) models include the neural network trained for feature selection using reinforcement learning (ML - RL), feature selection using Recursive Feature Elimination (ML - RFE), and a neural network trained with the full set of features (ML - NN). In addition, we present the performance of the initial rule-based model (RB 1) from interviews with therapists and that of the fine-tuned rule-based model (RB 10) after accommodating therapist's feature-based feedback. For hybrid models (HMs), we also described the performance of the HM 1 that combines ML-RL with RB 1 and that of the HM 10 that combines ML-RL with RB 10.

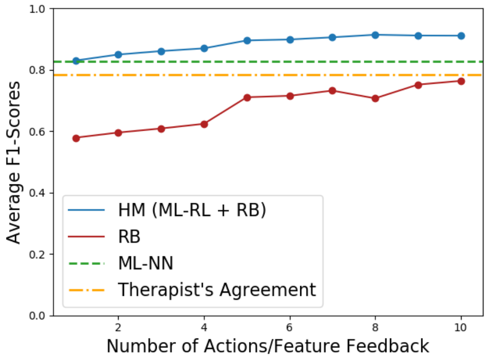
For feature selection, our approach has 0.11 higher average F1-score ($p < 0.01$ using a paired t-test over 3 exercises and 3 performance components) than a model with the Recursive Feature Elimination (RFE) approach (Guyon et al., 2002), one of classical feature selection methods, and is expected to perform better to generate patient-specific analysis for therapists.

Although a machine learning model with reinforcement learning-based feature selection (ML-RL) has slightly lower performance than a machine learning model with neural networks (ML-NN), ML-RL still achieves a decent agreement level with therapist's assessment: 0.8119 average F1-scores over three exercises. In contrast, the initial rule-based model (RB 1) achieves the lowest agreement level with therapist's assessment: 0.5827 average F1-scores over three exercises. The non-interactive, initial hybrid model (HM 1) integrates ML-RL with RB 1 achieves 0.8305 average F1-scores over three exercises, which is comparable with the performance of the ML-NN.

5.2. Effect of Therapists' Feature-based Feedback

For the evaluation of our interactive approach, therapists reviewed user-specific analysis of our system, and provided nine feature-based feedback on each patient to tune a system. On average, therapists added 7.26 new features, removed 0.33 features, and updated 1.06 threshold values over 15 patients.

While accommodating therapist's additional nine feature-based feedback on each patient, both rule-based (RB) model and hybrid model (HM) significantly improve their performance on all exercises (Figure 3). The RB model improves its performance 31% from 0.5821 to 0.7642 average F1-scores over all exercises ($p < 0.01$ using a paired t-test over 3 exercises and 3 performance components). Similarly, our interactive Hybrid Model (HM) also significantly improves its performance 9.7% from 0.8305 to 0.9116 average F1-scores ($p < 0.01$), which outperforms the ML with Neural Networks (ML - NN) and therapist's agreement level (TPA in Table 1).



(a)

Figure 3. The performance of models while accomodating therapist's feature-based feedback

6. Discussion

Our results demonstrate how an imperfect machine intelligence can complement with human intelligence on a complex task (e.g. rehabilitation assessment). As initial high-level rules from therapists are not tuned for each individual patient, the initial rule-based model (RB 1 in Table 1) performs worse than machine learning (ML) models. This implies the necessity of generating personalized rules to assess the performance of patients with various conditions.

Our approach can automatically identify salient features for assessment and generate a succinct user-specific analysis to elicit therapist's feedback for personalized rehabilitation assessment. After better understanding patient's exercise performance with quantitative data, therapists can provide feature-based feedback to refine an imperfect model. While accommodating therapist's feature-based feedback, the rule-based (RB) model improves its performance significantly, which is better than a machine learning model with recursive feature elimination (ML-RFE) and close to the therapist's agreement level (i.e. TPA in Table 1). This tuned RB model provides another valuable perspective on assessment, which also leads to the improvement on the performance of the Hybrid Model (HM). Specifically, our interactive HM (i.e. HM 10 in Table 1) achieves significantly higher performance than therapist's agreement level (i.e. TPA in Table 1). This implies the feasibility of consistently replicating therapist's assessment to improve the current practices of monitoring patient's exercises.

Overall, the results of our interactive hybrid approach demonstrate how machine intelligence and rule-based models from human intelligence can complement each other for a more accurate, personalized rehabilitation assessment model. However, although feature selection supports to generate a patient-specific analysis, as a supplementary explanation on a machine learning model, the full interpretation of a model still remains challenge to elicit expert's feedback. In addition, this work explores only feature-based feedback during few interactions with therapists. Further study is required to investigate whether this approach can also be personalized over a longer time period while patient's functional ability changes, and can be applied to another exercises and tasks.

7. Conclusion

In this paper, we present an interactive hybrid approach that can automatically generate a user-specific analysis to support therapist's understanding on patient's performance and accommodate therapist's feedback for more accurate, personalized rehabilitation assessment. Our experimental results show that after reviewing a patient-specific analysis, therapists can provide feedback to tune an imperfect model to a personalized model with improved performance. Our work highlights the importance of an interactive approach that can complement an imperfect machine learning model with expert's knowledge.

A. Appendix

Table 2. Parameters of Neural Networks

	Hidden Layers and Units / Learning Rate		
	ROM	Smooth	Comp
E1	(32, 32, 32) / 0.1	(16) / 0.0001	(256, 256) / 0.1
E2	(256) / 0.1	(512, 512) / 0.1	(128) / 0.1
E3	(256) / 0.1	(64, 64) / 0.001	(128, 128) / 0.1

Acknowledgements

This work is partially supported by the FCT (SFRH/BD/113694/2015, LARSyS - Plurianual funding 2020-2023, and CMU-PT IntelligentCare project) and the National Science Foundation (NSF) under grant number CNS-1518865.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 3. ACM, 2019.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182, 2016.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2):423–443, 2019.

Das, S., Trutoiu, L., Murai, A., Alcindor, D., Oh, M., De la Torre, F., and Hodgins, J. Quantitative measurement of motor symptoms in parkinson's disease: A study with full-body motion capture data. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6789–6792. IEEE, 2011.

Dulac-Arnold, G., Denoyer, L., Preux, P., and Gallinari, P. Datum-wise classification: a sequential approach to sparsity. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 375–390. Springer, 2011.

Fails, J. A. and Olsen Jr, D. R. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pp. 39–45. ACM, 2003.

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

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

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

Janisch, J., Pevný, T., and Lisý, V. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3959–3966, 2019.

Jones, M., Grimmer, K., Edwards, I., Higgs, J., and Trede, F. Challenges in applying best evidence to physiotherapy. *Internet Journal of Allied Health Sciences and Practice*, 4(3):11, 2006.

Khairat, S., Marc, D., Crosby, W., and Al Sanousi, A. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics*, 6 (2):e24, 2018.

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

Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pp. 126–137, 2015.

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernadino, A., et al. Learning to assess the quality of stroke rehabilitation exercises. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 218–228. ACM, 2019.

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., and 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, 2020a.

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., and 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, 2020b.

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., and i Badia, S. B. Opportunities of a machine learning-based decision support system for stroke rehabilitation assessment, 2020c.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529, 2015.

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

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

Patel, S., Hughes, R., Hester, T., Stein, J., Akay, M., Dy, J. G., and Bonato, P. A novel approach to monitor rehabilitation outcomes in stroke survivors using wearable technology. *Proceedings of the IEEE*, 98(3):450–461, 2010.

Saary, M. J. Radar plots: a useful way for presenting multivariate health care data. *Journal of clinical epidemiology*, 61(4):311–317, 2008.

Sanford, J., Moreland, J., Swanson, L. R., Stratford, P. W., and Gowland, C. Reliability of the fugl-meyer assessment for testing motor performance in patients following stroke. *Physical therapy*, 73(7):447–454, 1993.

Siewiorek, D., Smailagic, A., and Dey, A. Architecture and applications of virtual coaches. *Proceedings of the IEEE*, 100(8):2472–2488, 2012.

Tang, J., Alelyani, S., and Liu, H. Feature selection for classification: A review. *Data classification: Algorithms and applications*, pp. 37, 2014.

Taub, E., Morris, D. M., Crago, J., King, D. K., Bowman, M., Bryson, C., Bishop, S., Pearson, S., and Shaw, S. E. Wolf motor function test (wmft) manual. *Birmingham: University of Alabama, CI Therapy Research Group*, 2011.

Van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

Webster, D. and Celik, O. Systematic review of kinect applications in elderly care and stroke rehabilitation. *Journal of neuroengineering and rehabilitation*, 11(1):108, 2014.

Wu, C.-y., Trombly, C. A., Lin, K.-c., and Tickle-Degnen, L. A kinematic study of contextual effects on reaching performance in persons with and without stroke: influences of object availability. *Archives of Physical Medicine and Rehabilitation*, 81(1):95–101, 2000.

Xu, H., Gao, Y., Yu, F., and Darrell, T. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2174–2182, 2017.