Interactive Hybrid Approach to Combine Machine and Human Intelligence for Personalized Rehabilitation Assessment

Min Hun Lee
Daniel P. Siewiorek
Asim Smailagic
Carnegie Mellon University
{minhunl,dps,asim}@cs.cmu.edu

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

Sergi Bermúdez i Badia
Madeira Interactive Technology Institute
sergi.bermudez@m-iti.org

ABSTRACT

Automated assessment of rehabilitation exercises using machine learning has a potential to improve current rehabilitation practices. However, it is challenging to completely replicate therapist's decision making on the assessment of patients with various physical conditions. This paper describes an interactive machine learning approach that iteratively integrates a data-driven model with expert's knowledge to assess the quality of rehabilitation exercises. Among a large set of kinematic features of the exercise motions, our approach identifies the most salient features for assessment using reinforcement learning and generates a user-specific analysis to elicit feature relevance from a therapist for personalized rehabilitation assessment. While accommodating therapist's feedback on feature relevance, our approach can tune a generic assessment model into a personalized model. Specifically, our approach improves performance to predict assessment from 0.8279 to 0.9116 average F1-scores of three upper-limb rehabilitation exercises ($p < 0.01$). Our work demonstrates that machine learning models with feature selection can generate kinematic feature-based analysis as explanations on predictions of a model to elicit expert's knowledge of assessment, and how machine learning models can augment with expert's knowledge for personalized rehabilitation assessment.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → **Health care information systems**; • **Theory of computation** → *Sequential decision making*.

KEYWORDS

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

ACM Reference Format:
Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Interactive Hybrid Approach to Combine Machine and Human Intelligence for Personalized Rehabilitation Assessment. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20), April 2–4, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3368555.3384452

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7046-2/20/04…$15.00
https://doi.org/10.1145/3368555.3384452

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 [33]. 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 [33]. However, therapists have difficulty with making informed decision on adjusting treatment interventions without observing patient's exercises or quantitative exercise data [18].

Advanced sensor and machine learning technologies have a potential to automatically monitor and assess patient's status to support physical rehabilitation [42]. Previous work on computer-assisted rehabilitation monitoring and assessment can be categorized into rule-based and data-driven approaches [38]. A rule-based approach derives a set of monitoring rules through the involvement of experts in the design process [16]. However, it is difficult to properly articulate an expert's decision making process on a complex monitoring task.

Alternatively, a data-driven model utilizes machine learning algorithms with labeled sensor data to automatically extract a meaningful function (e.g. Neural Network model) for assessing the quality of motion [6, 28, 35]. However, it is challenging to derive a model that can replicate therapist's assessment for every patient, given that each patient has 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 [20, 22].

In this paper, we describe and evaluate an interactive hybrid approach that integrates a data-driven model with expert's knowledge on kinematic features to assess the quality of motion (Figure 1). Our approach 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 [28]. From this dataset, we apply reinforcement learning to identify the most important features for assessment and learn a machine learning

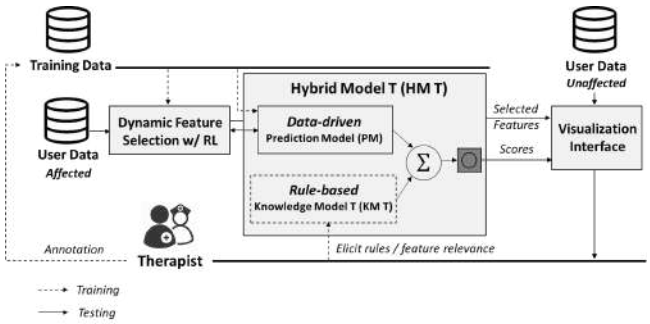


Figure 1: Flow diagram of an interactive hybrid approach for rehabilitation assessment that combines a data-driven model and a rule-based knowledge model. Data-driven models automatically select features and predict the quality of motion to generate a patient-specific report. A therapist can review this report and provide feature-based feedback to tune a model for personalized rehabilitation assessment

model to predict the scores on the exercises of patients using leave-one-subject-out cross validation. For the development of the initial rule-based Knowledge Model (KM), we conducted semi-structured interviews with therapists to elicit their knowledge of assessing rehabilitation exercises. The data-driven Prediction Model and rule-based Knowledge Model are integrated with a weighted average ensemble technique [4] to derive the Hybrid Model (HM) for 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. Therapists can review this analysis to better understand patient's performance and provide feedback (e.g. 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 therapist's feature-based feedback to significantly improve performance of a model from 0.8377 to 0.9116 average F1-scores on three exercises ($p < 0.01$).

This paper makes the following contributions:

- present an interactive hybrid, machine learning approach that can identify salient features for assessment and present a user-specific analysis to iteratively accommodate expert's feedback for personalized rehabilitation
- describe the evaluation of our approach with therapists to explore whether therapists can accurately personalize the predictions of machine learning-based rehabilitation assessment

Although prior work demonstrated the feasibility of monitoring and assessing rehabilitation exercises [42], there is a lack of evaluations on how such technologies can be utilized by therapists. To the best of our knowledge, this paper describes the first evaluation on how data-driven machine learning systems can augment with therapist's feedback for personalized rehabilitation assessment.

2 RELATED WORK

Researchers have explored the feasibility of monitoring and assessing chronic diseases with computational models [42], which can assist therapists to obtain insights on patient's status.

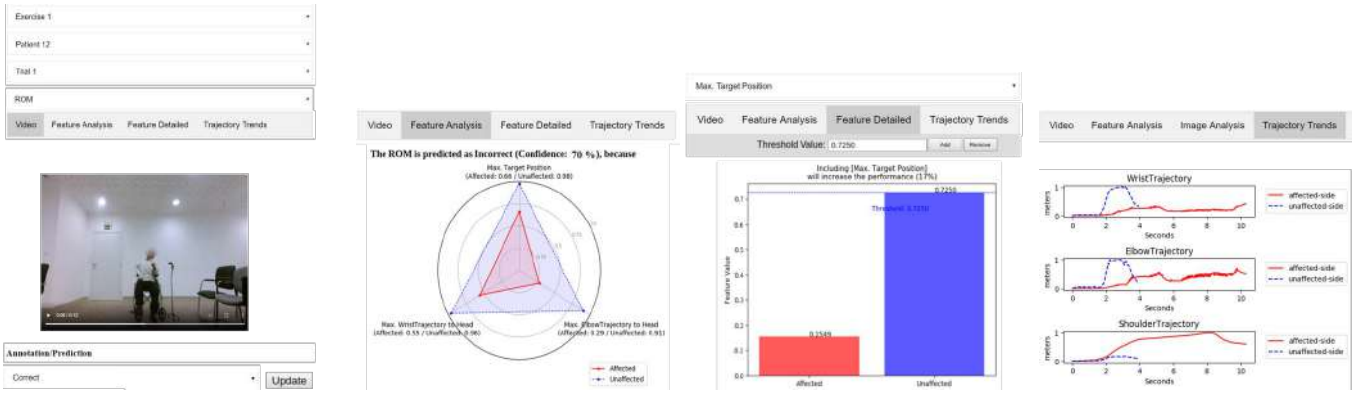
One approach is to elicit a set of monitoring rules from domain experts [38]. For instance, Huang explored an authoring tool that enables a therapist to specify repetitions and joint angles for monitoring knee rehabilitation exercises [16]. This rule-based approach can be easily modularized and recombined to develop a customized monitoring model. However, it is time consuming to manually determine which sensor measurement could be utilized to monitor an individual status. Moreover, experts might have difficulty with articulating their complex and abstract decision making process into a set of rules.

Another approach is to utilize machine learning with labeled sensor data [38] and automatically learn a meaningful model (e.g. Neural Networks) to assess the quality of motion [6, 28, 35]. However, as patients have various physical functional abilities, it is challenging to perfectly replicate therapist's assessment. When a model with complex algorithms cannot explain its prediction to support expert's decision making [13], it can exacerbate therapist's user experience and trust, which impedes its adoption in practices [20, 22].

Explainability [5, 7, 8] and interactive machine learning [1, 10, 15] have been an active area of research to create better machine learning models with improved transparency and user acceptance. Prior work on interactive machine learning has demonstrated the feasibility to refine the classification of a system while presenting relevant information of a task and acquiring inputs of a user. For instance, users can provide feedback on constraints of a model [19], weights of features [24], or feature relevance [23, 31] to tailor the behavior of a model.

Our work aims to increase the interpretability on a prediction model by feature selection [5, 21]. Specifically, we apply reinforcement learning [41] to identify salient features for assessment. Utilizing an identified subset of features, we predict the quality of motion and generate a user-specific analysis to summarize patient's exercise performance. This user-specific analysis will be presented to therapists in a visualization interface to improve their understanding on patient's performance [29].

In addition, this paper validates that our approach can present a user-specific analysis and elicit expert's feedback on feature relevance to iteratively update a model and improve performance on predictions for personalized rehabilitation assessment [27]. While interactive machine learning techniques have been applied to various tasks (e.g. text classification [23] and image search [12]), the application on a health domain seems to be elusive [15]. This work contributes to increase the knowledge on how machine learning



(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.

models can iteratively augment with therapist's feedback for personalized rehabilitation assessment.

3 STROKE REHABILITATION AS A TEST DOMAIN

Stroke is the second leading cause of death and third most common contributor to disability [11]. Thus, we selected stroke rehabilitation as a probe domain. We recruited three therapists of stroke rehabilitation from two rehabilitation centers to specify the design of our study (i.e. exercises and performance components for assessment).

After having iterative discussion with three therapists (TP 1, TP2, and TP3 with $\mu = 6.33$, $\sigma = 2.05$ years of experience in stroke rehabilitation) in Table 1, we specified the design of our study to assess stroke patient's rehabilitation exercises.

3.1 Three Task-Oriented Upper Limb Exercises

This paper utilizes three upper-limb stroke rehabilitation exercises (Figure 3), recommended by therapists [25]. In Figure 3, the *'Initial'* label indicates the initial position of an exercise and the *'Target'* label describes the desired end position of an exercise.

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. These exercises are selected due to their correspondence with major motion patterns: elbow flexion for Exercise 1, shoulder flexion for Exercise 2, elbow extension for Exercise 3.

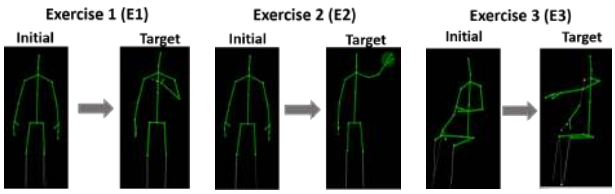
3.2 Performance Components

After reviewing commonly used stroke assessment tools (i.e. Fugl Meyer Assessment [37] and Wolf Motor Function Test [40]) and having iterative discussion with therapists, three common performance components and their scoring guidelines for therapists are specified to assess the quality of motion: *'Range of Motion (ROM)'*, *'Smoothness'*, and *'Compensation'* (Table 2). For binary labels, 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 *'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

Table 1: The participants of the specification, of the annotation, of the rule elicitation (ElicitRule), and of the relabelling (Relabel), and of the feature elicitation (ElicitFeat)

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



(a) (b) (c)

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

Table 2: Guidelines to Assess Stroke Rehabilitation Exercises

Performance Components	Score	Guidelines
Range of Motion (ROM)	0	Does not or barely involve any movement
	1	Less than half way aligned with an *'Target'* position
	2	Movement achieves an *'Target'* position
Smoothness	0	Excessive tremor or not smooth coordination
	1	Movement influenced by tremor
	2	Smoothly coordinated movement
Compensation	0	Noticeable compensation in more than two joints
	1	Noticeable compensation in a joint
	2	Does not involve any compensations

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

Figure 4: Two patients performing Exercise 1 with different compensated motions: the patient on the left has elevated shoulder with trunk rotation and the patient on the right has elevated shoulder with leaning backward.

shoulder to raise the affected hand as shown in Figure 4b and 4d. Each patient might have different compensated movements based on patient's functional status: one patient has elevated shoulder with trunk rotation (Figure 4b) and the other patient has elevated shoulder with leaning backward (Figure 4d).

3.3 Features

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

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) [28]. 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 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 [28]. 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 ($x \in R^{5d}$). In summary, the *'ROM'* has 30 features, the *'Smoothness'* has 60 features, and the *'Compensation'* has 65 features.

4 INTERACTIVE HYBRID APPROACH FOR REHABILITATION ASSESSMENT

Automated machine learning approaches make great progress in various fields that can afford a large dataset (e.g. speech recognition [3] and autonomous vehicles [44]). However, the healthcare domain often involves a small dataset, which makes the application of automated approaches difficult or even impossible [15]. Interactive machine learning approach seems to be a promising approach while making use of human cognitive abilities [10, 23, 26].

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

4.1 Dynamic Feature Selection using Reinforcement Learning

Kinematic feature analysis is an important source for therapists to quantitatively and objectively understand patient's performance [43]. 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 Prediction Model (PM) and generate an interpretable and succinct patient-specific report.

The classical approaches of feature selection (e.g. filter, wrapper, embedded methods) [39] find a fixed feature set to the entire training dataset for all patients. In contrast, this paper applies a Markov Decision Process (MDP) [17] to find the optimal feature set for each patient's motions. As each patient has different physical and functional status (Figure 4), we hypothesize that feature selection with MDP can perform better than classical feature selection approaches for personalized rehabilitation assessment.

4.1.1 Problem Definition. We formulate 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 [32, 41].

We mathematically describe the Markov Decision Process (MDP) with similar notations of [9, 17] 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 $\lambda = 0.01$. The agent receives a reward of -1 for incorrect classification and a reward of 0 for correct classification.
- **Transition**: Let the transition function be
$$p(s, a) = p((x, y, \bar{\mathcal{F}}), a) = \begin{cases} (x, y, \mathcal{F} \cup a) & \text{if } a \in \mathcal{A}_f \\ TS & \text{if } a \in \mathcal{A}_c \end{cases},$$
where TS is the terminal state after outputting the classification and revealing the true label.

4.1.2 Implementation Details. We utilize *'PyTorch'* libraries [34] to implement a neural network with parameters θ (Q_θ) for deep Q-learning [32]. The input layer of the network consists of feature and binary mask vectors [17]. 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 the Prediction Model (PM). The architectures and parameters of the PM are described in the Table 5.

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 [17] and update the target network after each step. Instead of directly updating the weight of the target network, we apply soft target updates [30]: $\theta' \leftarrow \rho\theta + (1 - \rho)\theta'$, 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 [32] for sampling efficiency. Specifically, the environment with randomly drawn samples is simulated and the transition data is recorded to the experience replay buffer. As the environment is episodic with a short length, we choose a value 1.0 for the discount factor γ. In addition, we apply the ϵ-greedy policy to control the exploration. Specifically, we linearly decrease the ϵ value from the $\epsilon_{start}(0.5)$ to the $\epsilon_{end}(0.05)$ with a step value, $\epsilon_{step}(0.02)$.

4.2 Prediction Model

The Prediction Model (PM) utilizes a supervised learning algorithm and training data from all patients except a patient for testing to predict the quality of motion or the posterior probability of being correct on a performance component, $P_{PM} = P(y = 1|x)$. We explore various traditional supervised learning algorithms: Decision Trees (DTs), Linear Regression (LR), Support Vector Machine (SVM), Neural Networks (NNs) using the *'Scikit-learn'* [36] and the *'PyTorch'* libraries [34].

For DTs, we apply Classification and Regression Trees (CART) to build prune trees. For LR models, we apply $L1, L2$ regularization or linear combination of $L1$ and $L2$ (ElasticNet with 0.5 ratio) to avoid overfitting. For SVMs, we apply either linear or Radial Basis Function (RBF) kernels with penalty parameter, $C = 1.0$. 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.

4.3 Knowledge Model

The Knowledge Model (KM) utilizes the set of feature-based rules from therapists to estimate the quality of a motion. For the initial development of the KM, we conducted semi-structured interviews 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*.

Let us denote a joint position as $p(j, c)$, where j specifies a joint (e.g. $shoulder(sh)$, $wrist(wr)$), c denotes a coordinate of joints in the set $C \in \{c_x, c_y, c_z\}$. For example, the assessment on the ROM component for Exercise 1 is specified as follows:

$$PredictedLabel = \begin{cases} \text{'Correct'} & \text{if } p^{max}(wr, c_y) >= p^{max}(sh, c_y) \\ \text{'Incorrect'} & \text{else} \end{cases}$$

This rule simply checks the maximum position of a wrist joint, $p^{max}(wr, c_y)$, related to that of a shoulder joint, $p^{max}(sh, c_y)$, in y-coordinate to roughly estimate whether a patient achieves a target position of Exercise 1 (Figure 6a). The score of being correct on each performance component using KM can be computed using the following equation:

$$P_{KM^T} = \frac{1}{|\mathbb{R}^T|} \sum_{s \in R^T} \min(\frac{f_s}{\tau_s}, 1) \quad (2)$$

where f_s indicates the feature value of a rule s from a trial (e.g. $p^{max}(wr, c_y)$ for the example above), τ_s describes the threshold value of a rule s (e.g. $p^{max}(sh, c_y)$ for the example above). $\mathbb{R}^T$ describes the set of rules considered relevant by 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 likelihood of being correct.

Furthermore, the KM 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 [23] 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 of a feature-based rule (τ_s).

4.4 Hybrid Model

The Hybrid Model (HM) applies a weighted average, ensemble technique [4] to integrate two perspectives on assessment: data-driven, Prediction Model (PM) and rule-based, Knowledge Model (KM) from therapists.

For the classification of the quality of motion, the 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, given a test sample (x, y), we compute the prediction score of HM, P_{HM} as follows:

$$P_{HM} = \frac{\rho_{pm}}{\rho_{pm} + \rho_{km^T}} P_{PM} + \frac{\rho_{km^T}}{\rho_{pm} + \rho_{km^T}} P_{KM^T} \quad (3)$$

where P_{PM} and P_{KM^T} indicate the scores of Prediction Model (PM) and Knowledge Model (KM) at the T iteration respectively, and ρ_{pm} and ρ_{km^T} describe the F1-scores of PM and KM^T.

4.5 Visualization Interface

Based on the prior work that describes the needs of therapists during stroke rehabilitation assessment [29] and the guidelines of Human Artificial Intelligence (AI) interaction [2, 23], 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 discussion with therapists from five rehabilitation centers, therapists 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 [29]. To present *"contextually relevant information"* [2] for the assessment, this interface presents a video of patient's exercise motion along with a user-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"* [2], 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 [29], this interface follows this current practice, *"social norms"* [2], 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"* [23] 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 [29].

In addition, our interface supports to *"honor user feedback"* (e.g. feature-based feedback) [23]. 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).

Prior work conducted an user study with therapists to evaluate the values of such an interface with predicted assessment and user-specific analysis [29]. It shows that the interface with predicted quality of motion and feature analysis improves therapist's understanding on patient's performance. Specifically, our approach with predicted assessment and user-specific analysis assists therapists to achieve significantly higher agreement level on evaluation (0.71 F1-score) than the baseline interface without prediction and user-specific analysis (0.66 F1-score) ($p < 0.05$) [29]. Prior work [29] confirms that such an interface supports more consistent assessment, and is preferred over the baseline interface without prediction and analysis. The details on the evaluation of the interface is described in [29].

5 EXPERIMENT FOR IMPLEMENTATION

5.1 Exercise Dataset

After the approval of ethics committee, we collected the dataset of three upper limb exercises from 15 post-stroke and 11 healthy subjects using a Kinect v2 sensor (Microsoft, Redmond, USA).

For the data collection, we implemented a program that records the 3D trajectory of joints and video frames at 30 Hz. The program was operated on a PC with 8GB RAM and i5-4590 3.3GHz 4 Cores CPU and a sensor was located at a height of 0.72m above the floor and 2.5m away from a subject. The starting and ending frames of exercise movements were manually annotated during the data collection.

All subjects signed the consent form before participating in the data collection. 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) [37]. 15 stroke survivors have diverse functional abilities from mild to severe impairment (37 ± 21 Fugl Meyer Scores). During the second session, a stroke survivor 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 as follows:

- ***'Training Data'*** (Figure 1) is composed of 165 unaffected motions from 11 healthy and 150 affected motions from 15 post-stroke subjects to train a feature selection model and the Prediction Model (PM).
- ***'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 data-driven 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 user-specific analysis of the visualization interface (Figure 2).

Two therapists (TP 1 and 2 in Table 1) 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. During the annotation, they had no access to analysis of our system (Figure 2d, 2c, 2d). For implementation, we utilize the annotation of therapist 1 (TP 1), who had more interactions with recruited stroke patients by supporting the recruitment and evaluation on their functional ability with Fugl Meyer Assessment. The annotation of therapist 2 (TP 2) is compared with that of TP1 to measure therapist's agreement on F1-scores in Table 3.

5.2 User 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 (i.e. TPs with check marks in the *'ElicitFeat'* column of Table 1) and analyzed the effect of therapist's feedback on the system performance for predicting rehabilitation assessment. In addition, we conducted semi-structured interviews with therapists to understand a potential benefit of our approach in their workflow.

After signing the IRB approved consent form, each participant 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 participant 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 participant to generate feature-based feedback on all post-stroke survivors in our dataset. Given each patient's affected motions, our approach dynamically selects salient features for assessing corresponding motions to predict the quality of motion and generate patient-specific analysis (Figure 1). After reviewing an assigned patient's user-specific analysis, participants provided nine feature-based feedback on each patient. Lastly, we interviewed therapists about the possibility of accepting our approach in the current practices.

6 RESULTS

6.1 Implementation

To evaluate the implementation of our approach, we apply Leave-One-Subject-Out (LOSO) cross validation on post-stroke patients. 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, a F1-score is utilized as therapists considered that less false positives (reporting normal when it is abnormal) and false negatives (reporting abnormal when it is normal) are crucial.

For feature selection, we train a reinforcement learning model with Neural Networks that sequentially decides whether another feature is necessary to assess the quality of motion, which is described in the Section 4.1. We illustrate the learning curve of training a reinforcement learning agent for feature selection by plotting the average rewards and the average number of selected features. Figure 5 demonstrates that an agent can identify the salient subset of features and reduce the number of selected features. At the same time, an agent improves its average rewards (i.e. the correct assessment of exercise motions).

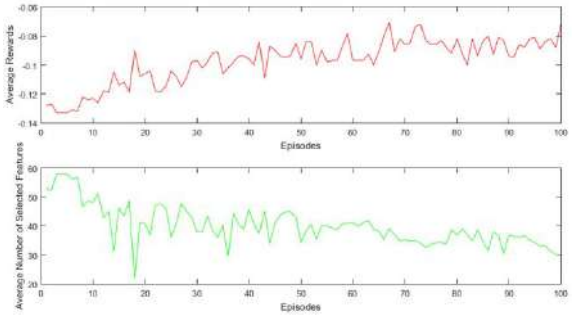
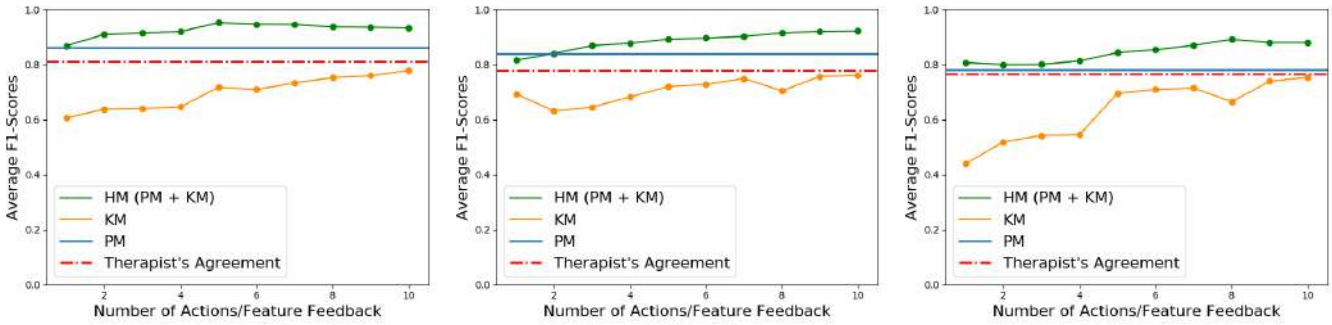


Figure 5: The average rewards and the average number of selected features while training an agent for dynamic feature selection

Table 3 summarizes the performance of our approach, which measures the agreement with therapist's evaluation using average F1-scores of the three exercises. The performance of models for individual performance components of each exercise (i.e. Range of Motion, Smoothness, and Compensation) are described in Table 4. The parameters of NNs (i.e. hidden layers/units and learning rate)



(a) Exercise 1 (b) Exercise 2 (c) Exercise 3

Figure 6: The performance of models over therapist's feature-based feedback: both Knowledge Model and Hybrid Model (HM) improves its agreement level while accommodating therapist's feature-based feedback. HM performs better than Prediction Model with Neural Network and therapist's agreement

Table 3: Performance (F1-scores) of interactive Hybrid Model (HM 10), non-interactive Hybrid Model (HM 1), Unimodal models (PM with various approaches, KM 1, KM 10), and Therapist's agreement

	Exercise 1 (E1)	Exercise 2 (E2)	Exercise 3 (E3)	Overall
PM - RL	0.8331± 0.0059	0.7973 ± 0.0867	0.8053 ± 0.0496	0.8119 ± 0.0526
PM - RFE	0.6742 ± 0.0715	0.7628 ± 0.1708	0.6415 ± 0.0806	0.6928 ± 0.1147
PM - DT	0.6901 ± 0.0405	0.7645 ± 0.0867	0.6488 ± 0.0412	0.7011 ± 0.0769
PM - LR	0.7246 ± 0.0593	0.6430 ± 0.0982	0.7267 ± 0.0391	0.6981 ± 0.0801
PM - SVM	0.7232 ± 0.0364	0.6971 ± 0.0891	0.7410 ± 0.0052	0.7204 ± 0.0585
PM - NN	0.8632 ± 0.0816	0.8388 ± 0.0518	0.7818 ± 0.0096	0.8279 ± 0.0605
KM 1	0.6148 ± 0.1702	0.6932 ± 0.1630	0.4384 ± 0.1569	0.5821 ± 0.1066
KM 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**
Therapist's Agreement	0.8120 ± 0.1458	0.7790 ± 0.1324	0.7654 ± 0.1382	0.7854 ± 0.0195

that achieve the best F1-score on the classification are summarized in the Table 5.

For Prediction Models (PMs), we evaluated various approaches: the neural network trained for feature selection using reinforcement learning (PM - RL), feature selection using Recursive Feature Elimination (PM - RFE) [14], a decision tree (PM-DT), a linear regression model (PM - LR), a support vector machine (PM - SVM), a neural network trained with the full set of features (PM - NN).

In addition, we presented the performance of the initial Knowledge Models (KM 1) from the interviews with therapists and that of the fine-tuned Knowledge Models (KM 10) after accommodating therapist's feature-based feedback. For the Hybrid Models (HMs), we also described the performance of the HM 1 without accommodating additional therapist's feedback and that of the HM 10 with additional therapist's feedback.

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 in Table 4) than a model with the Recursive Feature Elimination (RFE) approach, one of classical feature selection methods, and is expected to perform better to generate patient-specific analysis for therapists.

The data-driven, Prediction Model with Neural Networks (PM - NN) achieves a decent agreement level with Therapist 1's evaluation: 0.8279 average F1-scores over three exercises. In addition, the PM with NNs outperforms the PM with other algorithms: Decision Trees (0.7011 average F1-scores), Linear Regression (0.6981 average F1-scores), Support Vector Machine (0.7204 average F1-scores).

In contrast, the initial rule-based, Knowledge Model 1 (KM 1) achieves low agreement level with Therapist 1's evaluation: 0.5827 average F1-scores over all exercises. Non-interactive, Hybrid Model 1 (HM 1) that integrates PM with reinforcement learning and KM 1 achieves 0.8305 average F1-scores over three performance components of three exercises, which is slightly higher overall performance than the PM with Neural Networks (i.e. PM - NN in Table 3). However, integrating two modalities of assessment does not always improve the performance of a model (e.g. the performance of the HM 1 for Exercise 2).

6.2 User Study with Therapists

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. Specifically, on average, they 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 Knowledge Model (KM) and Hybrid Model (HM) improve their performance on all exercises (Figure 6). The KM improves its agreement level 31% from 0.5821 to 0.7642 average F1-scores over all exercises. Similarly, our interactive Hybrid Model (HM) also significantly improves it agreement level 9.7% from 0.8305 to 0.9116 average F1-scores ($p < 0.01$), which outperforms the PM with Neural Networks (PM - NN) and therapist's agreement (Table 3). Specifically, the improvement is statistically significant using paired t-test over 3 exercises and 3 performance components (Table 4).

Table 4: Performance comparison between the proposed method and baseline models to assess the quality of motion (*'Correct'* and *'Incorrect'*) with F1-Scores. Best results have been boldfaced. ‡ indicates HM10 is statistically better than the compared method (pairwise t-tests at 99% significance level)

Models	Algorithms	Exercise 1			Exercise 2			Exercise 3		
		ROM	Smooth	Comp	ROM	Smooth	Comp	ROM	Smooth	Comp
PM	RL ‡	0.8265 ± 0.2826	0.8351 ± 0.2586	0.8378 ± 0.2360	0.7998 ± 0.3442	0.7093 ± 0.3119	0.8828 ± 0.2210	0.7482 ± 0.2775	0.8372 ± 0.2373	0.8307 ± 0.2372
	RFE ‡	0.7013 ± 0.4316	0.5931 ± 0.4521	0.7284 ± 0.3658	0.9161 ± 0.1893	0.5787 ± 0.4789	0.7938 ± 0.3430	0.5726 ± 0.3942	0.7302 ± 0.3771	0.6217 ± 0.3706
	DT ‡	0.6366 ± 0.4339	0.7349 ± 0.3794	0.6988 ± 0.3969	0.8630 ± 0.2683	0.7786 ± 0.3513	0.6520 ± 0.3667	0.7062 ± 0.3967	0.6109 ± 0.3494	0.6295 ± 0.3886
	LR ‡	0.7899 ± 0.3396	0.7377 ± 0.2670	0.6463 ± 0.3633	0.7679 ± 0.3487	0.6335 ± 0.3683	0.5277 ± 0.4288	0.7728 ± 0.3061	0.7302 ± 0.3771	0.6771 ± 0.3914
	SVM ‡	0.7276 ± 0.3860	0.7655 ± 0.3388	0.6765 ± 0.3910	0.8028 ± 0.3449	0.7038 ± 0.3364	0.5848 ± 0.4485	0.7471 ± 0.3433	0.7342 ± 0.2988	0.7419 ± 0.3543
	NN ‡	0.9527 ± 0.0942	0.7927 ± 0.3424	0.8443 ± 0.2641	0.8367 ± 0.3363	0.7881 ± 0.2988	0.8917 ± 0.1284	0.7775 ± 0.3378	0.7751 ± 0.3255	0.7929 ± 0.2105
KM 1	Therapist ‡	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
KM 10	Therapist ‡	0.8994 ± 0.2510	0.6956 ± 0.4177	0.6289 ± 0.4189	0.8799 ± 0.3933	0.6120 ± 0.4851	0.6187 ± 0.4746	0.6689 ± 0.4683	0.7302 ± 0.3771	0.6656 ± 0.4931
HM 1	WeightedAvg‡	0.9321 ± 0.2041	0.8204 ± 0.3009	0.7987 ± 0.3409	0.8558 ± 0.3369	0.6855 ± 0.3794	0.6145 ± 0.4338	0.8011 ± 0.3202	0.6720 ± 0.3584	0.7603 ± 0.3818
HM 10	**WeightedAvg**	**0.9867 ± 0.0381**	**0.9280 ± 0.1051**	**0.9439 ± 0.1135**	**0.9926 ± 0.0277**	**0.9647 ± 0.0793**	**0.9733 ± 0.0444**	**0.9828 ± 0.0444**	**0.9113 ± 0.1203**	**0.9617 ± 0.00647**
Therapist's Agreement ‡		0.9587 ± 0.0489	0.5490 ± 0.0011	0.7289 ± 0.0298	0.9630 ± 0.0427	0.6588 ± 0.1384	0.8223 ± 0.0004	0.7342 ± 0.2418	0.5373 ± 0.1148	0.9046 ± 0.0497

According to the interviews with therapists on the possibility of accepting our approach in their workflow, therapists mentioned that our approach with predicted assessment and user-specific analysis is *"easy to use"*, and *"quickly summarizes quantitative data to provide useful insights on patient's performance"* [29]. Although the prediction of our approach sometimes mismatched with therapist's assessment, therapists considered reviewing predicted assessment and user-specific analysis of our approach is useful to *"validate whether a system fails to predict correctly or I make a mistake"* [29]. Overall, therapists are positive to accept our approach in their practices.

7 DISCUSSION

Our results demonstrate how machine and human intelligence can work together for a complex task (e.g. rehabilitation assessment). As initial high-level rules from therapists are not tuned for each individual patient, the initial Knowledge Model (KM 1 in Table 3) performs worse than various approaches of data-driven Prediction Models (PMs). This implies the necessity of generating personalized rules to assess the performance of patients with various physical characteristics and functional abilities.

To elicit therapist's feedback for personalized rehabilitation assessment, our approach can automatically identify salient features for assessment and generate a succinct user-specific analysis. This user-specific analysis provides therapists new insights on patient's performance [29]. 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 KM improves its performance to a similar level of the PM with Support Vector Machine (i.e. PM - SVM in Table 3). This tuned KM 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 3) achieves significantly higher agreement with Therapist 1 than Therapist 2 (i.e. therapist's agreement in Table 3). This implies the feasibility of consistently replicating therapist's assessment to improve the current practice of monitoring patient's exercises and supporting therapist's decision making.

Overall, the results of our interactive hybrid approach demonstrate how data-driven and rule-based models can complement each other for a more accurate, personalized rehabilitation assessment model. However, although feature selection supports to generate a user-specific analysis, as a supplementary explanation on the prediction of a model, it is still challenging to derive a full interpretation of complex algorithms for accommodating user'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.

8 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 personalized rehabilitation assessment. Our experimental results show that presenting a user-specific analysis is useful to improve therapist's understanding on the task of rehabilitation assessment. Specifically, therapists can provide feedback to tune a generic model to a personalized model with improved performance. Our work highlights the importance of presenting a supplementary explanation on the prediction of a model and creating an interactive machine learning-based system that can augment data-driven models with expert's knowledge to tune imperfect models. We believe this study can be served as a valuable reference to develop an interactive machine learning-based system for critical, medical decision making tasks (e.g. rehabilitation assessment).

A APPENDIX

Table 5: 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

ACKNOWLEDGMENTS

The authors would like to appreciate all post-stroke and healthy subjects for their participation in the data collection, and the anonymous therapists for discussing the experimental designs and their participation in the experiments. Also, the authors would like to thank the anonymous reviewers for their valuable comments. This work was partially supported by the FCT through the Augmented Human Assistance project (CMUP-ERI/HCI/0046/2013) and SFRH/BD/113694/2015, and the LARSyS - FCT Plurianual funding 2020-2023.

REFERENCES

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

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

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*. 173–182.

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

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

[6] Samarjit Das, Laura Trutoiu, Akihiko Murai, Dunbar Alcindor, Michael Oh, Fernando De la Torre, and Jessica Hodgins. 2011. 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*. IEEE, 6789–6792.

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

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

[9] Gabriel Dulac-Arnold, Ludovic Denoyer, Philippe Preux, and Patrick Gallinari. 2011. Datum-wise classification: a sequential approach to sparsity. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 375–390.

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

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

[12] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. In *Proceedings of the sigchi conference on human factors in computing systems*. ACM, 29–38.

[13] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA)* 2 (2017).

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

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

[16] Kevin Huang. 2015. *Exploring in-home monitoring of rehabilitation and creating an authoring tool for physical therapists*. Ph.D. Dissertation. Carnegie Mellon University.

[17] Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. 2019. Classification with costly features using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3959–3966.

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

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

[20] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics* 6, 2 (2018), e24.

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

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

[23] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

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

[25] Min Hun Lee. 2018. A Technology for Computer-Assisted Stroke Rehabilitation. In *23rd International Conference on Intelligent User Interfaces*. 665–666.

[26] Min Hun Lee. 2019. Intelligent Agent for Assessing and Guiding Rehabilitation Exercises. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 6444–6445. https://doi.org/10.24963/ijcai.2019/911

[27] Min Hun Lee. 2019. An Intelligent Decision Support System for Stroke Rehabilitation Assessment. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 694–696.

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

[29] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Opportunities of a Machine Learning-based Decision Support System for Stroke Rehabilitation Assessment. arXiv:2002.12261

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

[31] Thorsten May, Andreas Bannach, James Davey, Tobias Ruppert, and Jörn Kohlhammer. 2011. Guiding feature subset selection with an interactive visualization. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 111–120.

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

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

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

[35] Shyamal Patel, Richard Hughes, Todd Hester, Joel Stein, Metin Akay, Jennifer G Dy, and Paolo Bonato. 2010. A novel approach to monitor rehabilitation outcomes in stroke survivors using wearable technology. *Proc. IEEE* 98, 3 (2010), 450–461.

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

[37] Julie Sanford, Julie Moreland, Laurie R Swanson, Paul W Stratford, and Carolyn Gowland. 1993. Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke. *Physical therapy* 73, 7 (1993), 447–454.

[38] Daniel Siewiorek, Asim Smailagic, and Anind Dey. 2012. Architecture and applications of virtual coaches. *Proc. IEEE* 100, 8 (2012), 2472–2488.

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

[40] Edward Taub, David M Morris, Jean Crago, Danna Kay King, Mary Bowman, Camille Bryson, Staci Bishop, Sonya Pearson, and Sharon E Shaw. 2011. Wolf motor function test (WMFT) manual. *Birmingham: University of Alabama, CI Therapy Research Group* (2011).

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

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

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

[44] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. 2017. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2174–2182.