Learning to Assess the Quality of Stroke Rehabilitation Exercises

Min Hun Lee
Daniel P. Siewiorek
Asim Smailagic
Carnegie Mellon University
mhlee@cmu.edu
{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

Due to the limited number of therapists, task-oriented exercises are often prescribed for post-stroke survivors as in-home rehabilitation. During in-home rehabilitation, a patient may become unmotivated or confused to comply prescriptions without the feedback of a therapist. To address this challenge, this paper proposes an automated method that can achieve not only qualitative, but also quantitative assessment of stroke rehabilitation exercises. Specifically, we explored a threshold model that utilizes the outputs of binary classifiers to quantify the correctness of a movements into a performance score. We collected movements of 11 healthy subjects and 15 post-stroke survivors using a Kinect sensor and ground truth scores from primary and secondary therapists. The proposed method achieves the following agreement with the primary therapist: 0.8254, 0.8091, and 0.7571 F1-scores on three task-oriented exercises. Experimental results show that our approach performs equally well or better than multi-class classification, regression, or the evaluation of the secondary therapist. Furthermore, we found a strong correlation ($R^2 = 0.95$) between the sum of computed exercise scores and the Fugl-Meyer Assessment scores, clinically validated motor impairment index of post-stroke survivors. Our results demonstrate a feasibility of automatically assessing stroke rehabilitation exercises with the decent agreement levels and clinical relevance.

CCS CONCEPTS

• **Computing methodologies** → **Intelligent agents**; **Activity recognition and understanding**; • **Applied computing** → **Health care information systems**.

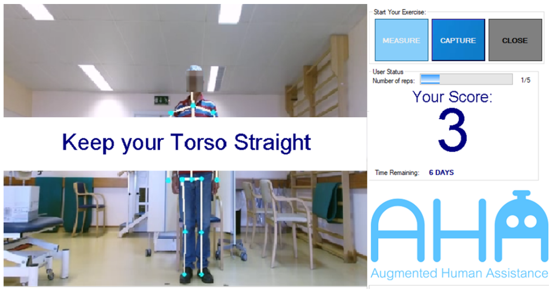
KEYWORDS

Intelligent Agent, Motion Analysis, Stroke Rehabilitation

ACM Reference Format:
Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2019. Learning to Assess the Quality of Stroke Rehabilitation Exercises. In *24th International Conference on Intelligent User Interfaces (IUI '19), March 17–20, 2019, Marina del Ray, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3301275.3302273

IUI '19, March 17–20, 2019, Marina del Ray, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302273



Number of reps: 1/5

Your Score: 3

Time Remaining: 6 DAYS

Figure 1: Computer-assisted stroke rehabilitation tool to assess the quality of patient's exercise using machine learning algorithms and a Kinect.

1 INTRODUCTION

One of the effective stroke rehabilitation approaches is physical therapy intervention with task-oriented exercises [21, 40, 47]. During a therapy session, a therapist monitors patient's performance and guides a patient with feedback. However, a post-stroke survivor may not receive timely and comprehensive rehabilitation due to the limited availability of therapists [37]. Alternatively, a therapist often prescribes in-home rehabilitation regimens, in which a post-stroke survivor independently participates in rehabilitation without any supervision of a therapist. A patient might feel uncertain whether he/she correctly performs an exercise and become unmotivated in rehabilitation regimens [2, 27]. A therapist encounters a challenge of tracking the progress of a patient and making an informative adjustment on rehabilitation regimens [15]. Thus, it becomes imperative to have a method to monitor and quantitatively assess in-home rehabilitation exercises.

Recent advances in sensors and machine learning algorithms offer a potential of computer-based in-home rehabilitation [20]. Researchers have demonstrated systems that log measurements of movements (e.g. joint angles [4]), count repetitions of exercises, and detect incorrect movements [16, 19, 31] to support in-home rehabilitation. However, logging motion-related measurements and binary detection of incorrect movements are not intuitive and sufficient to follow the progress of a post-stroke survivor's in-home rehabilitation. A therapist mentioned it is difficult to interpret sensor measurements (e.g. velocity of a joint) [16]. In addition, reporting the results of binary detection on an incorrect movement is limited for both therapists and patients to track diverse degrees of functional abilities. Instead of these approaches, prior studies on in-home rehabilitation exercise tools with therapists and patients [7, 16] highlighted the importance and need of having quantitative performance results. Performance results can reinforce patient's motivation [9, 26, 39] and patient's adherence to rehabilitation regimens [7]. In addition, they are valuable for therapists to understand patient's performance and adjust regimens.

In this paper, we describe an approach of automatically assessing stroke rehabilitation exercises using a RGB depth sensor and machine learning algorithms. Our approach leverages a threshold model with binary classifications [14], providing the ability to assess performance qualitatively and quantitatively (Figure 1). Our insight is that assessing the quality of stroke rehabilitation exercises has the property of both classification and regression with categorical scales (e.g. *'0: cannot perform'*, *'1: partially perform'*, *'2: fully perform'* [38]). The state-of-the-art approaches to classify categorical scales assume that the categorical response is a measured latent continuous variable, which can be modeled with intervals on the real line [14, 23]. This assumption allows to learning a set of thresholds to divide data into categorical responses and quantifying the performance of a movement using a threshold model with binary classifications.

To demonstrate our proposed method, we recruited two therapists to specify the experimental designs (i.e. three performance components of task-oriented rehabilitation exercises, three upper-limb exercises). We then collected a dataset of three upper-limb exercises, which includes 900 motions from 15 post-stroke survivors, ground truth scores from primary and secondary therapists, and motor impaired scores of 15 post-stroke survivors with Fugl Meyer Assessment (FMA).

Using this dataset, we evaluate our approach of assessing the quality of a movement. First, we show that a threshold model with binary classifications performs better or equally good with multi-class classification or regression approaches using various machine learning algorithms (i.e. Decision Tree, Linear Regression, Support Vector Machine, Neural Network, Long Short Term Memory Network). It promises more scalable development of an automated assessment than multi-class or regression approaches that require expensive data collection from therapists and post-stroke survivors.

Second, our approach can achieve decent agreement levels with the primary therapist (0.83, 0.81, and 0.76 F1-scores on three exercises), which is equally good or better than the secondary therapist. In addition, the computed scores of our approach has a strong correlation with FMA scores. Our approach can consistently assess clinically relevant, quantitative performance scores without repetitively requiring multiple hours of discussion between therapists.

The contributions of this paper are as follows:

- We introduce an automated assessment method that utilizes a threshold model with binary classifications to qualitatively and quantitatively assess task-oriented rehabilitation exercises.
- We present the experimental results from 15 post-stroke survivors and two therapists on three upper-limb exercises across three performance components, which validate the effectiveness of our approach and clinical relevance with Fugl Meyer Assessment (FMA).

2 BACKGROUND AND RELATED WORK

2.1 Challenges of Stroke Rehabilitation Practices

Post-stroke survivors should receive an early and extensive rehabilitation program to prevent disability and stroke recurrence. Promoting task-oriented exercise is a popular strategy to improve functional ability and lower a chance of having recurrent stroke [40]. However, it is expensive and difficult for post-stroke survivors to receive the administration of an individualized rehabilitation session [37]. Instead, a post-stroke survivor engages in in-home rehabilitation without any supervision of a therapist.

Both stroke survivors and therapists encounter following challenges to pursue in-home rehabilitation. First, stroke survivors may have low participation in rehabilitation due to several reasons [2, 9]. Most of them expressed anxieties about a lack of information and support from professionals [2, 27]. They described the need of a trainer, who provides motivation and coaching on their performance [9]. Specifically, they mentioned that a way to see physical improvement after exercising would be a good source of exercise motivation [9]. According to the studies of [26, 39], they found that viewing their scores on a screen made them motivated to beat their previous scores. In addition, as patient's self-report is a primary source of a therapist to follow the adherence and progress of a patient, a therapist has limited quantitative performance data to understand patient's progress [15]. A therapist has difficulty with adjusting in-home rehabilitation regimens and deriving the general predictability of post-stroke motor recovery [15]. As a first step to address these challenges, this paper mainly explores the feasibility of developing a computer-assisted method to assess stroke rehabilitation exercises.

2.2 Computer-Assisted Rehabilitation Monitoring

One preliminary approach to monitor in-home rehabilitation is logging the measurements of a sensor. Huang utilized inertial sensors to measure joint movements (e.g. repetitions, rotation velocity, frequency, range of motion) for balance rehabilitation [16]. However, according to the user study of deploying a monitoring system [16], physical therapists mentioned that basic measurements (e.g. degrees/sec) were not intuitive to assess a patient's compliance to prescription. Physical therapists preferred to have easily comprehensible data, so that they could spend more time with patients [16].

Another approach is to apply gesture recognition [10, 25] that classifies incorrect movements. Chang et al. utilized the rules of six joint angles to identify the accuracy of movements for upper limb rehabilitation [6]. Pogorelc et al. utilized k-nearest neighbors and neural network algorithms to recognize four gait related problems with body-worn tags and wall-mounted sensors [36]. Das et al. applied a Support Vector Machine (SVM) classifier to distinguish mild and severe symptoms of Parkinson's Disease using full body motion capture data from four Parkinson's patients [10]. Su et al. computed joint positions and the speed of completing an exercise and applied neural networks and fuzzy logics to classify the quality of an exercise movement into three levels (e.g. Bad, Good, Excellent) [45].

This gesture-recognition approach provides the number of correct movements to motivate a patient's participation and follow a patient's progress. However, it has limitation to represent diverse levels of patient's progress. We cannot differentiate the performance of two patients, who have the equivalent number of incorrect movements with different degrees of incorrectness. Moreover, it would continuously indicate the incorrectness until patient's full recovery. It may de-motivate a patient at some point. Thus, this paper mainly focuses on exploring the feasibility of quantitatively assessing patient's exercise performance.

Several research works have shown the usefulness of kinematic variables to represent an objective assessment of upper limb motor performance. Using an acted-out dataset from healthy subjects, Zhao et al. described a potential benefit of joint angles and hand positions to evaluate patient's recovery [49]. Ozturk et al. demonstrated that the feasibility of using the speed and joint angle measurements to differentiate three stroke patients from two healthy subjects [30]. Murphy et al. identified that the measurement of compensatory trunk and arm movements can be utilized to discriminate moderate and mild arm impairment using *'reaching and drinking'* exercise [28]. Patterson et al. demonstrated that kinematic variables (e.g. peak velocity, trunk displacement, etc) may be feasible and useful to measure functional ability [33]. However, limited works address how these kinematic variables can be exploited to quantitatively assess post-stroke survivor's exercise performance and functional ability scores.

This leads to the following research questions:

- RQ1. How can we develop a system that automatically assesses exercise performance using kinematic variables?
- RQ2. How closely do computed scores of an automated method align with therapist's assessment?
- RQ3. Do computed scores of an automated method have any clinical relevance?

3 EXPERIMENTAL DESIGNS OF TECHNOLOGY PROB FOR ASSESSING EXERCISES

The goal of this paper is not to show comprehensive functionalities of a computer-assisted rehabilitation system. Instead, this paper focuses on presenting a method of assessing the quality of stroke rehabilitation to collect patient's performance data and further demonstrating how well it can assess compare to therapists as a technology probe [17, 18].

A therapist utilizes observation-based tests to assess the motor ability of a post-stroke survivor. We analyzed these existing functional assessment tests for stroke rehabilitation to identify therapist's assessment strategies. After having iterative discussion with therapists, we specified our experimental designs to demonstrate a feasibility of quantitatively assessing stroke rehabilitation exercises. In the following subsections, we describe functional assessment tools for stroke rehabilitation and our experimental designs (i.e. performance components, therapist's scoring guidelines, three upper limb exercises).

3.1 Functional Assessment Tests of Stroke Rehabilitation

The Fugl Meyer Assessment (FMA) [42] and the Wolf Motor Function Test (WMFT) [46] are frequently used to determine the motor ability of adult post-stroke survivors aged over 18 years old. The FMA examines the functional use of both upper and lower extremities through monitoring selected movement patterns. For each pattern, a therapist assigns the quality of movement on a 3-point ordinal scale (0 - 2): 0 = *'cannot preform'*, 1 = *'partially perform'*, and 2 = *'fully performs'*. The evaluation of upper extremity includes 33 tasks with the maximum of 66 points. The WMFT requires a therapist to measure time and assign a Functional Ability (FA) score to each of the 17 functional tasks on a 6-point ordinal scale (0-5). The scoring guidelines of the WMFT describe a single, combined statement that suggests considering the fluidity, precision of movements, and the existence of compensatory movements (e.g. the extent to which the head and trunk are maintained in normal alignment).

These clinical tests include a series of functional movements, where a therapist observes a targeted joint motion and assigns a numerical score to each functional movement using the scoring guidelines of a selected performance test. The summation of performance scores from functional movements represents post-stroke survivor's functional ability.

3.2 Performance Components of Task-Oriented Stroke Rehabilitation Exercises

We identified commonly used factors to assess a functional ability of stroke patients from popular stroke rehabilitation assessment tools (i.e. the FMA and the WMFT) and related work on computer-assisted stroke rehabilitation monitoring systems. After the discussion with therapists, we specify three performance components to provide more detailed assessment and feedback instead of having a single overall score.

Table 1 summarizes the descriptions of three performance components: *'Range of Motion (ROM)'*, *'Smoothness'*, and *'Compensation'*. This abstracted categorization of three performance components can be represented with various kinematic factors of a motion for stroke rehabilitation. We describe the definition of three performance components along with the citations of the related work, whose monitoring features can belong to our specified performance components, and therapist's scoring guidelines.

One primary performance component is to evaluate whether a post-stroke survivor can achieve a particular motion pattern or task. The *'ROM'* component represents the amount of an active movement with a specific joint. The *'ROM'* component can be represented by joint angles [6, 16, 30, 42, 49, 50] or joint trajectory positions [5, 45, 46, 49].

Another performance component is to check the existence of jerky motion patterns or compensated movements. The *'Smoothness'* component represents the degree of trembling and irregular movement patterns of joints. The *'Smoothness'* component can be indicated by velocity related features [5, 33, 41, 42, 46].

The *'Compensation'* component monitors if a post-stroke survivor performs compensatory movements to achieve the target

Table 1: Three Performance Components and Therapist's Scoring Guidelines of Stroke Rehabilitation Exercises

Performance Components	Descriptions	Related Work	Binary Labels	Score	Therapist's Guidelines
Range of Motion (ROM)	The amount of movement around a specific joint	[5, 6, 16, 30, 42, 45, 46, 49, 50]	Incorrect	0	Does not or barely involve any movement
				1	Attempt is made but limited to be normal
			Correct	2	Movement appears to be normal
Smoothness	The degree of trembling movement	[5, 33, 41, 42, 46]	Incorrect	0	Excessive tremor or not smooth coordination
				1	Movement influenced by tremor
			Correct	2	Smoothly coordinated movement
Compensation	The involvement of compensation to perform a motion	[5, 33, 46]	Incorrect	0	Noticeable compensation in more than two joints
				1	Noticeable compensation in a joint
			Correct	2	Does not involve any compensations

postures. Specifically, during upper limb movements, therapists focus on the occurrence of the following compensatory movements: leaning or swaying torso movements or elevated shoulder joints [5, 33, 46] as shown in Figure 2b.

Figure 2: Motions of (a) unaffected side and (b) affected side with compensated shoulder and trunk joints due to limited functional ability

3.3 Three Upper Limb Exercises

This paper utilizes three upper limb stroke rehabilitation exercises (Figure 3) to validate the proposed approach. These exercises represent examples of task-oriented movements for stroke rehabilitation. In Figure 3, the *'Initial'* indicates the initial position of an exercise and the *'Target'* describes the desired position of an exercise. Exercise 1 is *'Bring a Cup to the Mouth'*, in which a subject has to hold a cup and raise it to the mouth as if drinking water. Exercise 2 is *'Switch a Light On'*, where a patient pretends touching a light switch on the wall with shoulder forward flexion movement. Exercise 3 is *'Move a Cane Forward'*, which aims to practicing the usage of a cane while performing elbow extension at the seated position.

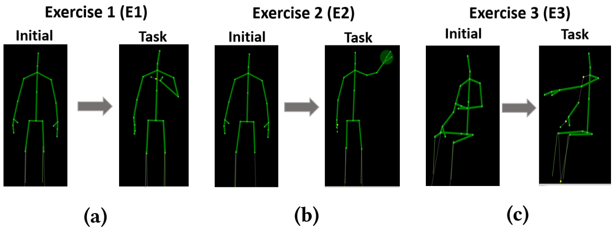


Figure 3: (a) Exercise 1 (b) Exercise 2 (c) Exercise 3

A therapist prescribes to repeat few exercises with major muscle groups during in-home rehabilitation [2, 34]. Based on the discussion with therapists, three exercise are selected due to their correspondence with major motion patterns: Exercise 1 to elbow flexion, Exercise 2 to shoulder flexion, and Exercise 3 to elbow extension [22].

In addition, previous studies have applied similar movement patterns to evaluate the usefulness of kinematic variables: *'reaching and drinking from a glass'* task [28], *'reaching'* task [33], and *'forward reaching'* movement [30]. These three exercises can be also mapped into movements in clinically validated assessment tests: the Fugl Meyer Assessment (FMA) and the Wolf Motor Function Test (WMFT). Exercise 1 is related to the *'Elbow Flexion'* movement of the FMA and the *'Lift can'* task of the WMFT. Exercise 2 is similar to the *'Shoulder Flexion'* movement of the FMA and the *'Hand to a box'* task of the WMFT. Exercise 3 is relevant to the *'Elbow Extension'* of the FMA and the *'Extend Elbow'* task of the WMFT.

Note that the purpose of this study is not to replace the existing functional ability tests but to demonstrate the feasibility of quantifying exercise performance score. This paper focuses on exploring the applicability of the proposed method to multiple patients using three exercises.

4 LEARNING A MODEL TO ASSESS THE QUALITY OF A MOVEMENT

An assessment of in-home rehabilitation exercises is to predict a categorical performance score (e.g. *'0: no movement perform'*, *'1: limited movement'*, *'2: normal movement'*), which can be considered as an intermediate problem between classification and regression. We can cast all categorical labels into real values and apply standard regression or multi-class classification techniques [3, 14]. One popular approach is called as a threshold model [14, 23], which assumes that categorical responses has a latent variable on the real line. This approach aims to learning a function, $f(\mathbf{x})$ that predicts the values of a latent variable [14] and estimating a categorical response with a set of threshold in the range of $f(\mathbf{x})$.

Prior work showed that well-tuned binary classification approaches can be transformed into good ranking algorithms [12, 23]. The confidence of a binary classifier can be considered as an ordering preference [13, 24]. Furthermore, Li and Lin showed theoretical and empirical analysis that this problem can be transformed into binary classifications [23]. Inspired by prior work, we utilize a threshold model with binary classifiers for assessing the quality of a movement.

The overall flow diagram of the proposed approach is described in Figure 4. Given an exercise trial, we extract various kinematic features. As a therapist assesses stroke rehabilitation exercise in terms of three performance components, we then train binary classifiers of performance components to predict latent variables. Leveraging these latent variables, we quantify a performance score with threshold models.

4.1 Feature Extraction

This section describes how we extract kinematic features and which features are utilized to model classifiers of performance components. This work applies a moving average filter with the window size of five frames to reduce noise of acquiring joint positions from a Kinect similar to [44].

This work uses the following notations to describe kinematic features. We denote a joint position as $p_t(j, c)$

- j specifies a joint in the set J, which includes selected tracking joints of a Kinect (Figure 6a).
- $J \in \{head(hd), spineshoulder(ss), shoulder(sh), elbow(eb), wrist(wr), spinemid(sm), spinebase(sb), hip(hp)\}$
- c denotes a coordinate of joints in the set $C \in \{x, y, z\}$.
- t is a frame number. T is the total number of frames.
- F is a sampling frequency, 30Hz

We process and normalize joint positions to reduce individual physical variabilities. The list of preprocessed and normalized features are described with their equations in Table 2. In addition to the common notations, we use a superscript to specify the statistics (i.e. max for the maximum and avg for the average). ft denotes a type of feature (i.e. sp for speed, ac for acceleration, and jk for jerk as defined in Table 2). $\mathbb{I}$ is an indicator function.

The joint angle (ja_t) feature computes an angle among three joints. The relative trajectory (rt_t) computes how far a selected joint is moved away from the basis joint using the Euclidean distance. We specify a head joint as the basis joint and elbow and wrist joints as selected joints for upper limb exercises. The projected trajectory (pt_t) describes the absolute distance between one selected joint and the other selected joint in a specific c coordinate. Using the relative trajectory features, we also compute the following quantities of a motion: speed (sp_t), acceleration (ac_t), and jerk (jk_t).

In addition, we define several normalized features to compensate individual's physical variability. The normalized relative trajectory (nrt_t) and normalized projected trajectory (npt_t) describe the change of a trajectory feature from an initial position. These normalized features can be considered as the normalization with subject's physical conditions, because an initial position is dependent on subject's physical characteristics (e.g. the length of limbs). It can utilize to segment a starting and ending frames of an exercise.

To normalize the speed related features, we utilize an average and maximum of speed or jerk until a selected frame. Normalized speed or jerk (nsp_t or njk_t) is the division of an average speed or jerk by the maximum speed or jerk value. If a subject has limited functional ability, a joint trajectory involves a number of valley shapes. His/her average speed or jerk with limited functional ability is expected to be smaller than that of health subjects. He/she is expected to have small value of normalized speed or jerk [41]. Mean Arrest Period Ratio (MAPR) represents the portion of frames when speed exceeds a percentage (10%) of the maximum speed [41]. As a subject reaches a target position without unnecessary stops, speed profiles will stay less near zero. We expect subjects with limited functional abilities will have more stationary movements and higher values of MAPR than health subjects. A zero-crossing ratio indicates the period of a movement, in which a sign of acceleration or jerk changes. Subjects with limited functional abilities will have stationary or trembling movements, which involve a number of valley shapes in speed profiles. Thus, post-stroke survivors with limited functional abilities expect to have higher zero-crossing ratio of acceleration or jerk than healthy subjects.

Table 2: List of Pre-processed and Normalized Features

Feature	Equaton
Joint Angle	$ja_t(j1, j2, j3) = \arccos(\frac{P_t(j1, j2) \cdot P_t(j2, j3)}{\lvert P_t(j1, j2)\rvert\lvert P_t(j2, j3)\rvert})$ $P_t(j1, j2) = (p_t(j1, x) - p_t(j2, x)) + (p_t(j1, y) - p_t(j2, y)) + (p_t(j1, z) - p_t(j2, z))$
Relative Trajectory	$rt_t(b, s) = \sqrt{\sum_{c \in C} (p_t(b, c) - p_t(s, c))^2}$
Projected Trajectory	$pt_t(j1, j2, c) = \lvert p_t(j1, c) - p_t(j2, c)\rvert$
Speed	$sp_t(j) = F * (rt_t(b, j) - rt_{(t-1)}(b, j)), \;\; if\, t > 1.$ $= 0\; otherwise$
Acceleration	$ac_t(j) = F * (sp_t(j) - sp_{(t-1)}(j)), \quad if\, t > 1.$ $= 0\; otherwise$
Jerk	$jk_t(j) = F * (ac_t(b, j) - ac_{t-1}(b, j)), \quad if\, t > 1.$ $= 0\; otherwise$
Normalized Relative Trajectory	$nrt_t(b, s) = \frac{\lvert rt_t(b,s) - rt_1(b,s)\rvert}{rt_1(b,s)}$
Normalized Projected Trajectory	$npt_t(j1, j2, c) = \frac{dpt_t(j1, j2, c)}{pt_1(j1, j2, c)}$ $dpt_t(j1, j2, c) = \lvert pt_t(j1, j2, c) - pt_1(j1, j2, c)\rvert$
Normalized Speed/Jerk	$nsp_t(j) = \frac{sp_t^{avg}(j)}{sp_t^{max}(j)}, \quad njk_t(j) = \frac{jk_t^{avg}(j)}{jk_t^{max}(j)}$
MAPR Speed/Jerk	$mapr_t(ft, j) = \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}_A(ft_s(j)),$ $A = \{ft_s(j) > ft_t^{max}(j) * 0.1\}, ft_s(j) \in \{sp_s(j), jk_s(j)\}$
Zero-Crossing Ratio	$zc_t(ft, j) = \frac{1}{(t-1)} \sum_{s=2}^{t} \mathbb{I}_{\mathbb{R}_{<0}}(ft_s(j) ft_{(s-1)}(j)),$ $ft_s(j) \in \{ac_s(j), jk_s(j)\} \quad \text{for } t > 1$

Given the list of preprocessed and normalized features, we describe the list of extracted features to train classifiers of performance components in Table 3.

The features of the *'ROM'* component include joint angles (ja_t), normalized relative trajectories (nrt_t), and projected trajectories (npt_t). The features of ja_t compute elbow and shoulder joint angles. The features of nrt_t measure the distance of wrist and elbow joints with respect to the head joint. The features of npt_t compute the distance of wrist with respect to the head and shoulder joints in x, y, z coordinates.

In addition, we represent the degree of smoothness with various speed-related features (i.e. speed (sp_t), acceleration (ac_t), jerk (jk_t), MAPR ($mapr_t$), zero-crossing ratios (zc_t), and normalized speed and jerk (nsp_t and njk_t) [41]). As this work includes only upper

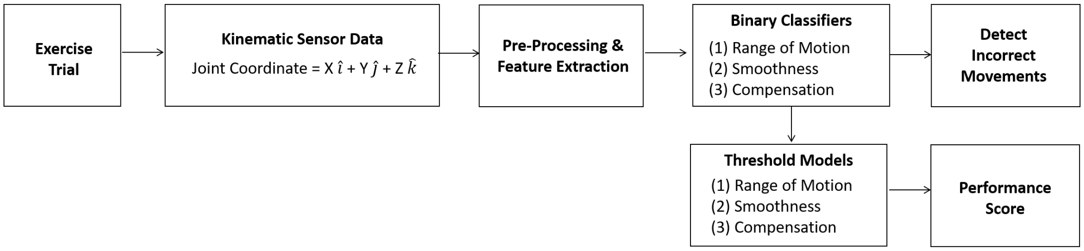


Figure 4: Flow diagram of the proposed method: Given an exercise trial, the system leverages kinematic sensor data to extract features. It trains binary classifiers on three performance components to detect the occurrence of incorrect movements and predict latent variables. Threshold models then utilize predicted latent variables to quantify a performance score.

Table 3: The list of extracted features for modeling classifiers of performance components

Performance Components	Features
ROM	$ja_t(hp, sh, eb), ja_t(sh, eb, wr)$ and $nrt_t(b, s)$ where $b = hd$ and $s \in \{eb, wr\}$ and $npt_t(j1, j2, c)$ where $(j1 = hd, j2 = wr)$ and $(j1 = sh, j2 = wr)$ for $c \in C$
Smoothness	$sp_t(j), ac_t(j), jk_t(j), nsp_t(j), njk_t(j),$ $mapr_t(sp, j), mapr_t(jk, j), zc_t(ac, j), zc_t(jk, j)$ where $j \in \{elbow, wrist\}$
Compensation	$ja_t(ss_{init}, sb_{init}, ss), ja_t(sh_{init}, ss_{init}, sh), ja_t(hp, sh, eb)$ $dpt_t(hd_{init}, hd, c), dpt_t(ss_{init}, ss, c), dpt_t(sh_{init}, sh, c)$ for $c \in C$

limb exercises, we extract those features on wrist and elbow joints for the *'Smoothness'* component.

For the *'Compensation'* component, we compute joint angles (ja_t) and projected trajectories (dpt_t) to distinguish a compensated movement. The features of ja_t calculate the tilted angle of a spine, the elevated angle of a shoulder, and shoulder abduction angle. The features of dpt_t measure the distance between the initial and current joint positions of head, spine, and shoulder joints in x, y, z coordinates.

4.2 Models to Assess the Quality of a Movement

This section describes our approach of developing models to assess the quality of a movement. We first apply standard (a) multi-class classification and (b) regression approaches as the baselines and then compare with the (c) *'BinToMulti'*, proposed approach, threshold models with binary classifications (Figure 4). We utilize *'Scikit-learn'* [35] and *'PyTorch'* [32] libraries to implement Decision Trees (DTs), Linear Regression (LR), Support Vector Machine (SVM), Neural Networks (NNs), and Long Short Term Memory (LSTM) Network.

For DTs, we utilize Classification and Regression Trees (CART) [1] to build a prune trees. For LR models, we apply either L1 or L2 regularization to avoid overfitting. For SVMs, we apply either linear or RBF kernals with penalty parameter, C = 1.0 using Support Vector Classification/Regression (SVC/SVR). For NNs, we explore various architectures (i.e. one to three layers with 16, 32, 64, 128, 256, 512 hidden units) and with adaptive learning rate with various initial learning rates (i.e. 0.005, 0.001, 0.01, 0.1). To explore the usefulness of applying a sequential model, we implement LSTM networks. As a therapist assesses the quality of a motion after observing patient's entire motion, LSTM networks have many-to-one architecture (Figure 5). We apply 0.5 drop-out to LSTM layers, explore one to three LSTM layers with 128, 256, 512 hidden units, and also apply three fully connected layers (the same hidden unit size with LSTM layers) to generate an output. For training NNs and LSTM networks, we apply *'ReLu'* activation functions, *'AdamOptimizer'* with mini-batch size of 5 and epoch = 1. Cross-entropy loss is utilized for classification and mean-square-error loss for regression approach.

We train classifiers of individual performance components using the features in Table 3. For each exercise trial, we compute five statistics (i.e. max, min, range, average, and standard deviation) of features at each time-stamp. Sequential models utilize the feature matrix over the entire time-stamps. Non-sequential models (i.e. DTs, LRs, SVMs, and NNs) apply feature vectors at the last time-stamp that summarizes the entire motion with statistics of features.

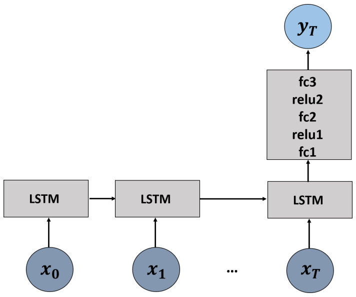


Figure 5: Many-to-One Architecture of LSTM Network.

Our baseline multi-class classification and regression approaches train classifiers of performance components with multi-class labels. In contrast, our proposed approach first trains classifiers of performance components with binary labels. These binary classifiers can predict the correctness of performance components and generate a confidence score to estimate performance score [13, 14, 24, 48] as shown in Figure 4.

Our threshold model can model the confidence of binary classifications with a real-valued function, $f : \mathcal{X} \times \{0, 1, .., K-1\} \rightarrow \mathbb{R}$ and estimate a performance score as follows:
Let denote the input vector $\mathbf{x} \in \mathcal{X}$ and label, $y \in \mathcal{Y} = \{0, 1, .., K-1\}$.

$f_b(\mathbf{x})$ describes the output, confidence score of binary classifications.

$$s(\mathbf{x}) \overset{\text{def}}{=} \sum_{i=1}^{K-1} [\![f(\mathbf{x}, i) > 0]\!] \qquad (1)$$

where $f(\mathbf{x}, i) = f_b(\mathbf{x}) - \theta_i$, $\theta_i = \frac{i}{K}$, $[\![a]\!]$ is defined to be 1 if a holds and 0 otherwise.

5 VALIDATING MODELS TO ASSESS THE QUALITY OF A MOVEMENT

This section describes our experiment to validate models to assess the quality of a movement using the data set from 11 healthy subjects, 15 post-stroke survivors, and two therapists. First, we demonstrate how well the assessment of models can be aligned with a therapist. Second, we show that quantified performance scores of proposed assessment method can have strong correlation with the scores of Fugl-Meyer Assessment, clinically validated motor impairment tool.

5.1 Data Collection

We collected an exercise movement dataset using a Kinect v2 sensor (Microsoft, Redmond, USA). The data collection program is implemented in C# using Kinect SDK and Accord.NET framework [43] and operated on a PC with 8GB RAM and i5-4590 3.3GHz 4 Cores CPU. This program records the trajectory of the selected body joints in Figure 6a and captures video frames at 30 Hz. The sensor was located at a height of 0.72m above the floor and 2.5m away from the subject (Figure 6b).

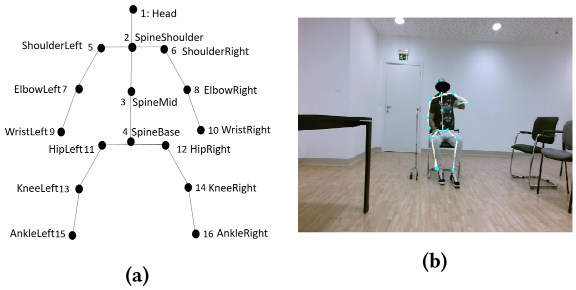


Figure 6: (a) Selected joints of a Kinect 2 (b) An exemplary captured frame of data collection

5.1.1 Participants.

Both healthy subjects and post-stroke survivors contributed to the dataset of three upper limb exercises (Figure 3) after signing the consent form. 11 healthy subjects (10 males and 1 female) with an average and standard deviation of 35.3 ± 5.81 years were recruited to collect unaffected movements. Healthy subjects were instructed to perform 15 repetitions for each exercise.

In addition, 15 post-stroke survivors (13 males and 2 females) with an average and standard deviation age of 63 ± 11.43 years participated in the data collection. A post-stroke survivor performed 10 trials with both affected and unaffected sides for each exercise. Thus, we have 465 trials for each upper limb exercise: 315 trials of unaffected movements and 150 trials of affected movements. The starting and ending frames of exercise movements are manually annotated for the experiment. In addition to recording exercise motions, a therapist assessed post-stroke survivor's functional ability of using the Fugl Meyer Assessment (FMA). The profiles of 15 post-stroke participants are summarized in Table 4.

Table 4: Profiles of 15 post-stroke survivors

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

5.1.2 Ground Truth Scores.

We recruited two stroke rehabilitation therapists: the primary and secondary therapists. The primary therapist managed the recruitment of participants in Table 4 and evaluated their functional ability with Fugl Meyer Assessment (FMA). The secondary therapist has an experience in stroke rehabilitation, but no prior interactions with recruited participants.

Two therapists conducted two evaluation phases to generate ground truth scores (Figure 7). In the first phase, each therapist individually watched the videos of participant's motions and assigned scores of each performance components using the scoring guidelines (Table 1). In the second phase, two therapists discussed their evaluation strategies for an hour by using one sample trial and then individually assigned scores again.

5.2 Level of Agreement with Therapist's Ground Truth

As a post-stroke survivor mainly interacts with a single therapist for the consistent assessment and guidance [11, 40], our study mainly explores how well our proposed method can be calibrated with the primary therapist's assessment and applicable to multiple patients.

The overall procedure to measure agreement levels with the primary therapist is described in Figure 7. We first measured the agreement between primary and secondary therapists during two evaluation phases, which can provide an idea of how well secondary therapist can align without or with one-hour discussion. We then computed the agreement levels between primary therapist and computer-assisted approaches (i.e. proposed and alternative approaches). Even if alternative approaches (i.e. multi-class classification and regression) are not shown in Figure 7, we also computed their agreement levels using numerical ground truth of primary therapist's 2nd evaluation. We then compared agreement levels of the proposed approach with alternative approaches and secondary therapist to analyze any benefits of the proposed method.

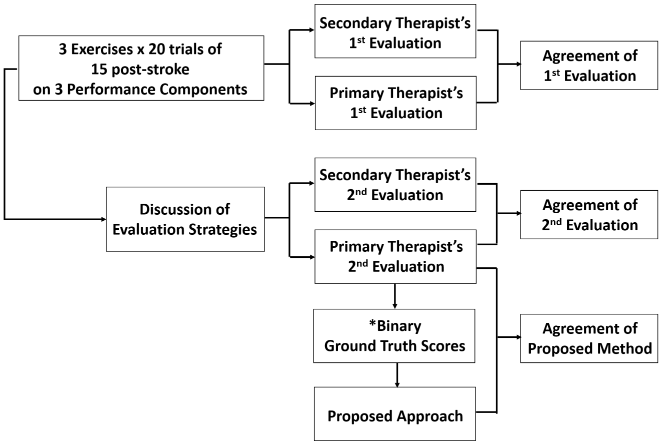


Figure 7: Procedure to Compute Agreement with Primary Therapist

For the evaluation of computer-assisted approaches, we apply Leave-One-Subject-Out (LOSO) cross validation on post-stroke survivors, which trains a model with data from all subjects except one post-stroke survivor and test with data from the left-out post-stroke survivor.

According to the experiments of binary classification (Table 5), Decision Trees (DTs) outperform other algorithms for *'E2-ROM'* (max-depth=3) and *'E3-Smooth'* (max-depth=4) and Neural Networks (NNs) outperform others algorithms for the rest of exercise performance components: *'E1-ROM'* with hidden layers (128, 128) and initial learning rate (1e-3), *'E1-Smooth'* with hidden layers (128, 128, 128) and initial learning rate (1e-3), *'E1-Comp'* with hidden layers (512, 512) with initial learning rate (1e-1), *'E2-Smooth'* with hidden layers (512, 512) and initial learning rate (1e-1), *'E2-Comp'* with hidden layers (256, 256, 256) and initial learning rate (1e-1), *'E3-ROM'* with hidden layers (128, 128) and initial learning rate (1e-1), *'E3-Comp'* with hidden layers (256, 256, 256) and initial learning rate (1e-3)).

Using these algorithms and parameters, we present the agreement levels of computer-assisted approaches (i.e. *'Multi-Class'* classification, *'Regression'*, and the proposed, *'BinToMulti'*) in Table 6. The highest F1-scores and lowest MSE of computer-assisted approaches are highlighted in a bold font. In addition, Table 6 presents the agreement levels of secondary therapist's evaluation, in which the highest F-scores and lowest MSE are highlighted in an italics font.

Overall, *'BinToMulti'* achieves the following average F1-scores: E1 with 0.8254, E2 with 0.8091, E3 with 0.7571. Table 6 shows that *'BinToMulti'* has higher F1-scores and lower MSE values than other computer-assisted approaches or secondary therapist's evaluation on some performance components. According to the one sample t-test with the results of Table 6, *'BinToMulti'* is equally good with other computer-assisted approaches on F1-scores and MSE values ($p < 0.01$). The same trends still hold after including performances of other algorithms (i.e. LRs, SVMs, LSTMs).

Compared to secondary therapist's 1st and 2nd evaluation, *'BinToMulti'* has significantly higher F1-scores on *'Smoothness'* and *'Compensation'* components of three exercises ($p < 0.05$) and equally good on *'ROM'* ($p < 0.01$). It has equally good MSE values on three components ($p < 0.05$).

Moreover, we include the scattered performance ((1-F1 scores) on y-axis and MSE on x-axis) of computer-assisted approaches with various algorithms and secondary therapist (Figure 9). We are able to identify *'BinToMulti'* or other computer-assisted approaches that performs better (located toward an origin) than secondary therapist except for *'E2-ROM'*.

5.3 Relationship with Fugl Meyer Scores

We analyze the relationship between the sum of computed scores using the proposed method and scores of the Fugl Meyer Assessment (FMA) using a linear regression. The linear model has 0.95 R-squared value ($p < 0.001$), which indicates a strong linear relationship between computed scores of the proposed method and the FMA scores of 15 post-stroke survivors (Figure 8). Using this linear regression model and the computed scores of exercises, our approach can estimate the Fugl Meyer score of a post-stroke survivor.

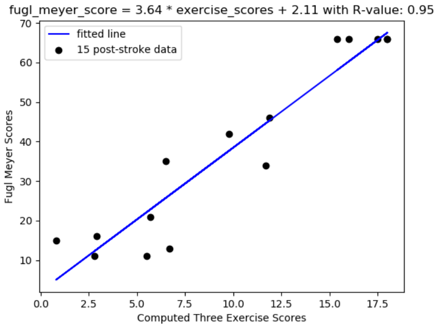


Figure 8: Regression between computed scores of proposed method and Fugl Meyer scores (R-value = 0.95, $p < 0.001$).

6 DISCUSSION

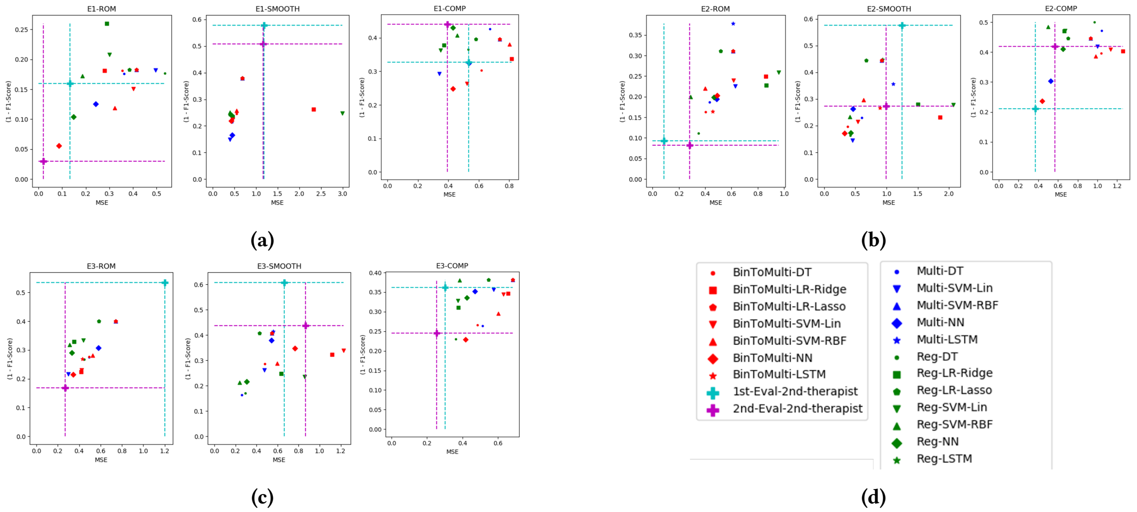
Our approach provides a method to automatically assess stroke rehabilitation using kinematic features and machine learning algorithms. Several studies have demonstrated the usefulness of kinematic variables to describe motion functional abilities [28, 30, 33, 49]. Furthermore, Olesh et al. described the feasibility of measuring the movement impairment with linear regression between two principal components of four joint angles and qualitative scores from eight subjects [29]. However, their study is limited to individual joint analysis and not applicable for complex task-oriented exercises. The current study involves the experimental designs for assessing task-oriented exercises (i.e. *'ROM'*, *'Smoothness'* *'Compensation'* performance components) with more kinematic variables and demonstrates the effectiveness of the proposed automated assessment.

Table 5: Binary Classification Results

	Exercise 1						Exercise 2						Exercise 3					
Elements	**ROM**		**Smoothness**		**Compensation**		**ROM**		**Smoothness**		**Compensation**		**ROM**		**Smoothness**		**Compensation**	
Algorithms	f1	mse	f1	mse	f1	mse	f1	mse	f1	mse	f1	mse	f1	mse	f1	mse	f1	mse
DT	0.8901 ± 0.2069	0.1067 ± 0.1922	0.8907 ± 0.1772	0.1133 ± 0.1658	0.7533 ± 0.2665	0.2133 ± 0.2093	**0.8793** **± 0.2572**	**0.1202** **± 0.2528**	0.9089 ± 0.1196	0.1009 ± 0.1222	0.6971 ± 0.2683	0.2733 ± 0.2461	0.7653 ± 0.2778	0.2044 ± 0.2267	**0.9043** **± 0.1833**	**0.1067** **± 0.1759**	0.7671 ± 0.2377	0.2111 ± 0.1878
LR	0.8689 ± 0.2187	0.0959 ± 0.0882	0.8702 ± 0.1709	0.5250 ± 1.5607	0.6909 ± 0.2748	0.2178 ± 0.1998	0.8038 ± 0.2892	0.2109 ± 0.4432	0.8921 ± 0.1355	0.4270 ± 0.8142	0.6737 ± 0.2648	0.2553 ± 0.1754	0.6000 ± 0.3266	0.2278 ± 0.1525	0.5922 ± 0.3150	0.2300 ± 0.1463	0.6186 ± 0.3161	0.2198 ± 0.1526
SVM	0.8173 ± 0.2924	0.1367 ± 0.2194	0.6206 ± 0.3188	0.2833 ± 0.2392	0.6038 ± 0.3238	0.2967 ± 0.2425	0.6889 ± 0.3327	0.2333 ± 0.2496	0.5556 ± 0.3144	0.3333 ± 0.2359	0.5536 ± 0.3157	0.3351 ± 0.2370	0.6000 ± 0.3266	0.3000 ± 0.2449	0.5922 ± 0.3150	0.3049 ± 0.2366	0.6186 ± 0.3161	0.2844 ± 0.2365
NN	**0.9472** **± 0.1390**	**0.0467** **± 0.1118**	**0.9176** **± 0.1725**	**0.0767** **± 0.1365**	**0.8009** **± 0.2744**	**0.1800** **± 0.2249**	0.8627 ± 0.2888	0.1368 ± 0.2801	**0.9658** **± 0.0846**	**0.0468** **± 0.1245**	**0.7939** **± 0.2673**	**0.1809** **± 0.2169**	**0.7928** **± 0.2625**	**0.1814** **± 0.2143**	0.7192 ± 0.2722	0.2537 ± 0.2107	**0.8297** **± 0.1768**	**0.1646** **± 0.1601**
LSTM	0.8173 ± 0.2924	0.1367 ± 0.2194	0.6206 ± 0.3188	0.2833 ± 0.2392	0.7033 ± 0.2425	0.2967 ± 0.2425	0.6889 ± 0.3327	0.2333 ± 0.2496	0.8395 ± 0.2138	0.2051 ± 0.2515	0.5536 ± 0.3157	0.3351 ± 0.2370	0.6000 ± 0.3266	0.3000 ± 0.2449	0.5922 ± 0.3150	0.3049 ± 0.2366	0.6186 ± 0.3161	0.2844 ± 0.2365

Table 6: Comparison of Agreement Level among 1) Multi-Class Classification, 2) Regression, 3) Proposed Approach, BinToMulti, 4) 1st-evaluation and 5) 2nd-evaluation of secondary therapist

		F1-Score					MSE				
		Multi-Class	Regression	**BinToMulti**	First_Eval	Second_Eval	Multi-Class	Regression	**BinToMulti**	First_Eval	Second_Eval
Exercise 1 (E1)	ROM	0.8746 ± 0.2180	0.8962 ± 0.1712	**0.9442** **± 0.1019**	0.8400	*0.9700*	0.2433 ± 0.4936	0.1484 ± 0.2038	**0.0867** **± 0.1384**	0.1333	*0.0200*
	Smoothness	**0.8351** **± 0.2119**	0.7593 ± 0.2084	0.7807 ± 0.2182	0.4219	*0.4913*	0.4500 ± 0.5908	0.4268 ± 0.3905	**0.4267** **± 0.4785**	1.1866	*1.1666*
	Compensation	0.6758 ± 0.3135	0.5689 ± 0.3038	**0.7515** **± 0.3085**	*0.6725*	0.5594	0.5367 ± 0.6744	0.4305 ± 0.3685	**0.4300** **± 0.5495**	0.5333	*0.3933*
Exercise 2 (E2)	ROM	0.8133 ± 0.2944	**0.8886** **± 0.2488**	0.8367 ± 0.2874	0.9076	*0.9177*	0.4321 ± 0.9703	**0.3495** **± 0.9865**	0.4033 ± 0.9802	*0.0872*	0.2818
	Smoothness	0.7376 ± 0.2608	0.8260 ± 0.1752	**0.8276** **± 0.2136**	0.4241	*0.7270*	0.4586 ± 0.5392	0.4226 ± 0.6342	**0.3240** **± 0.4948**	1.2483	*0.9932*
	Compensation	0.6961 ± 0.2924	0.5896 ± 0.3133	**0.7630** **± 0.2894**	*0.7896*	0.5813	0.5274 ± 0.6195	0.6504 ± 0.6332	**0.4402** **± 0.5744**	*0.3691*	0.5704
Exercise 3 (E3)	ROM	0.6930 ± 0.2871	0.7101 ± 0.2935	**0.7854** **± 0.2405**	0.4656	*0.8318*	0.5819 ± 0.6117	**0.3311** **± 0.3360**	0.3451 ± 0.4013	1.2013	*0.2684*
	Smoothness	**0.8362** **± 0.1812**	0.8293 ± 0.1816	0.7143 ± 0.2361	0.3929	*0.5622*	**0.2600** **± 0.3200**	0.2925 ± 0.3767	0.4782 ± 0.5372	*0.6644*	0.8657
	Compensation	0.6478 ± 0.2748	0.6643 ± 0.2499	**0.7715** **± 0.2322**	0.6385	*0.7549*	0.4711 ± 0.3816	0.4240 ± 0.4125	**0.4179** **± 0.5070**	0.3020	*0.2550*



(a)

(b)

(c)

(d)

Figure 9: Scatter Plots of (1-F1 Scores) and MSE, in which an origin (0, 0) indicates perfect agreement with ground truth scores. Multi-Class (Blue), Regression (Green), BinToMulti (Red), 1st-Evaluation (Cyan) and 2nd-Evaluation (Magenta) of Secondary Therapist. (a) Exercise 1 (b) Exercise 3 (c) Exercise 3 (d) Legends and Markers of Scatter Plots.

We first empirically show that our proposed *'BinToMulti'* method can perform equally good or better than other approaches (i.e. multi-class classification and regression) to quantify the performance of a movement. This result implies that researchers can transform the problem of learning a quantification function into sub-problems, set of binary classification and utilize confidence scores of binary classification for better exploitation of samples and performance improvement. Our approach may be utilized as a way to address imbalance samples in a healthcare application (including computer-assisted stroke rehabilitation) [8], which has a costly process of collecting data samples due to widely distributed subjects. However, it is not safe to assume that this approach can be applicable to any data-sets. Additional theoretical and empirical analysis are necessary to achieve generalizability of this approach.

In addition, our approach is feasible to reproduce primary therapist's assessment with the decent levels of agreement (Table 6) and estimate a clinically validated functional ability score (Figure 8).

Even if the primary and secondary therapists engaged in one-hour discussion to share and develop common strategies of evaluation, this one-hour discussion is not sufficient to completely reduce therapist's subjective interpretation on the clinical functional assessment tool. One-hour discussion does not necessarily improve the agreement level on the *'Smoothness'* and *'Compensation'* components. For the further improvement on the agreement level between the primary and secondary therapists, they are required to conduct an additional expensive evaluation process, where they have to meet and analyze the disagreed trials one-by-one iteratively until the convergence.

Without having additional expensive evaluation process, our approach can achieve equally good and better agreement with the primary therapist than the secondary therapist without or with one hour discussion. Our proposed method can serve as a low-cost method that reproduces therapist's assessment with decent agreement to consistently collect patient's performance data with clinical relevance.

7 CONCLUSION AND FUTURE WORK

In this paper, we proposed a method that utilizes threshold models with binary classifications to assess performance of stroke rehabilitation exercises. This work contributed the empirical study that validates the effectiveness of the proposed approach. For the validation, we collected a dataset with 11 healthy subjects and 15 post-stroke survivors performing three task-oriented exercises and two therapists (i.e. primary and secondary) generating ground truth scores and assessing the functional ability scores of post-stroke survivors with Fugl Meyer Assessment (FMA). We empirically demonstrated that the proposed method can achieve equally good or better level of agreement with the primary therapist than other approaches (i.e. multi-class classification and regression) or the secondary therapist. In addition, we showed that our approach can estimate a clinical FMA score with the sum of computed scores.

Although this study shows a feasibility of the proposed method that can reproduce therapist's assessment with a decent agreement level (0.83, 0.81, and 0.76 F1-scores on three exercises) and assess the quality of a movement, it is still challenging to reach the perfect agreement level with computer-assisted approaches. According to our experiments, applying a complex model (i.e. LSTM models) does not necessary guarantee to improve the agreement levels with the therapist. In future, we would investigate a learning technique to derive a personalized assessment model and then compare it with an user-agnostic assessment model. Moreover, we would explore a way to generate explanations of an automated assessment method and derive a human-in-the-loop system, which can updates an assessment model based on feedback of a therapist to achieve a better agreement level and usability.

REFERENCES

[1] Richard A Berk. 2008. Classification and regression trees (CART). In *Statistical learning from a regression perspective*. Springer, 1–65.

[2] Sandra A Billinger, Ross Arena, Julie Bernhardt, Janice J Eng, Barry A Franklin, Cheryl Mortag Johnson, Marilyn MacKay-Lyons, Richard F Macko, Gillian E Mead, Elliot J Roth, et al. 2014. Physical activity and exercise recommendations for stroke survivors: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 45, 8 (2014), 2532–2553.

[3] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[4] Antonio Bo, Mitsuhiro Hayashibe, and Philippe Poignet. 2011. Joint angle estimation in rehabilitation with inertial sensors and its integration with Kinect. In *EMBC: Engineering in Medicine and Biology Conference*. 3479–3483.

[5] Helen Bourke-Taylor. 2003. Melbourne assessment of unilateral upper limb function: construct validity and correlation with the pediatric evaluation of disability inventory. *Developmental medicine and child neurology* 45, 2 (2003), 92–96.

[6] Yao-Jen Chang, Wen-Ying Han, and Yu-Chi Tsai. 2013. A Kinect-based upper limb rehabilitation system to assist people with cerebral palsy. *Research in developmental disabilities* 34, 11 (2013), 3654–3659.

[7] Roberto Colombo, Fabrizio Pisano, Alessandra Mazzone, Carmen Delconte, Silvestro Micera, M Chiara Carrozza, Paolo Dario, and Giuseppe Minuco. 2007. Design strategies to improve patient motivation during robot-aided rehabilitation. *Journal of neuroengineering and rehabilitation* 4, 1 (2007), 3.

[8] Ricardo Cruz, Kelwin Fernandes, Joaquim F Pinto Costa, María Pérez Ortiz, and Jaime S Cardoso. 2018. Binary ranking for ordinal class imbalance. *Pattern Analysis and Applications* (2018), 1–9.

[9] Teresa M Damush, Laurie Plue, Tamilyn Bakas, Arlene Schmid, and Linda S Williams. 2007. Barriers and facilitators to exercise among stroke survivors. *Rehabilitation nursing* 32, 6 (2007), 253–262.

[10] 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 *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 6789–6792.

[11] Pamela W Duncan, Richard Zorowitz, Barbara Bates, John Y Choi, Jonathan J Glasberg, Glenn D Graham, Richard C Katz, Kerri Lamberty, and Dean Reker. 2005. Management of adult stroke rehabilitation care: a clinical practice guideline. *stroke* 36, 9 (2005), e100–e143.

[12] Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *European Conference on Machine Learning*. Springer, 145–156.

[13] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research* 4, Nov (2003), 933–969.

[14] Pedro Antonio Gutierrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. 2016. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering* 28, 1 (2016), 127–146.

[15] Henk T Hendricks, Jacques van Limbeek, Alexander C Geurts, and Machiel J Zwarts. 2002. Motor recovery after stroke: a systematic review of the literature. *Archives of physical medicine and rehabilitation* 83, 11 (2002), 1629–1637.

[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] Kevin Huang, Patrick J Sparto, Sara Kiesler, Asim Smailagic, Jennifer Mankoff, and Dan Siewiorek. 2014. A technology probe of wearable in-home computer-assisted physical therapy. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2541–2550.

[18] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 17–24.

[19] Majid Janidarmian, Atena Roshan Fekr, Katarzyna Radecka, and Zeljko Zilic. 2014. Affordable erehabilitation monitoring platform. In *Humanitarian Technology Conference-(IHTC), 2014 IEEE Canada International*. IEEE, 1–6.

[20] Michelle J Johnson, Xin Feng, Laura M Johnson, and Jack M Winters. 2007. Potential of a suite of robot/computer-assisted motivating systems for personalized, home-based, stroke rehabilitation. *Journal of NeuroEngineering and Rehabilitation*

4, 1 (2007), 6.

[21] Peter Langhorne, Peter Sandercock, and Kameshwar Prasad. 2009. Evidence-based practice for stroke. *The Lancet Neurology* 8, 4 (2009), 308 – 309.

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

[23] Ling Li and Hsuan-Tien Lin. 2007. Ordinal regression by extended binary classification. In *Advances in neural information processing systems*. 865–872.

[24] Hsuan-Tien Lin and Ling Li. 2006. Large-margin thresholded ensembles for ordinal regression: Theory and practice. In *International Conference on Algorithmic Learning Theory*. Springer, 319–333.

[25] Roanna Lun and Wenbing Zhao. 2015. A survey of applications and human motion recognition with microsoft kinect. *International Journal of Pattern Recognition and Artificial Intelligence* 29, 05 (2015), 1555008.

[26] Grace A MacDonald, Nicola M Kayes, and Felicity Bright. 2013. Barriers and facilitators to engagement in rehabilitation for people with stroke: a review of the literature. *New Zealand Journal of Physiotherapy* 41, 3 (2013), 112–121.

[27] Niall Maclean, Pandora Pound, Charles Wolfe, and Anthony Rudd. 2000. Qualitative analysis of stroke patients' motivation for rehabilitation. *Bmj* 321, 7268 (2000), 1051–1054.

[28] Margit Alt Murphy, Carin Willén, and Katharina S Sunnerhagen. 2011. Kinematic variables quantifying upper-extremity performance after stroke during reaching and drinking from a glass. *Neurorehabilitation and neural repair* 25, 1 (2011), 71–80.

[29] Erienne V Olesh, Sergiy Yakovenko, and Valeriya Gritsenko. 2014. Automated assessment of upper extremity movement impairment due to stroke. *PloS one* 9, 8 (2014), e104487.

[30] Ali Ozturk, Ahmet Tartar, Burcu Ersoz Huseyinsinoglu, and Ahmet H Ertas. 2016. A clinically feasible kinematic assessment method of upper extremity motor function impairment after stroke. *Measurement* 80 (2016), 207–216.

[31] Jiann-I Pan, Hui-Wen Chung, and Jan-Jue Huang. 2013. Intelligent shoulder joint home-based self-rehabilitation monitoring system. *Int. J. Smart Home* 7, 5 (2013), 395–404.

[32] 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).

[33] Tara S Patterson, MD Bishop, TE McGuirk, A Sethi, and LG Richards. 2011. Reliability of upper extremity kinematics while performing different tasks in individuals with stroke. *Journal of motor behavior* 43, 2 (2011), 121–130.

[34] Bente Klarlund Pedersen and B Saltin. 2006. Evidence for prescribing exercise as therapy in chronic disease. *Scandinavian journal of medicine & science in sports* 16, S1 (2006), 3–63.

[35] 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.

[36] Bogdan Pogorelc, Zoran Bosnić, and Matjaž Gams. 2012. Automatic recognition of gait-related health problems in the elderly using machine learning. *Multimedia Tools and Applications* 58, 2 (2012), 333–354.

[37] Alexandra S Pollock, Lynn Legg, Peter Langhorne, and Cameron Sellars. 2000. Barriers to achieving evidence-based stroke rehabilitation. *Clinical Rehabilitation* 14, 6 (2000), 611–617.

[38] Melinda Randall, John B Carlin, Patty Chondros, and Dinah Reddihough. 2001. Reliability of the Melbourne assessment of unilateral upper limb function. *Developmental medicine and child neurology* 43, 11 (2001), 761–767.

[39] Denise Reid and Tasneem Hirji. 2004. The influence of a virtual reality leisure intervention program on the motivation of older adult stroke survivors: A pilot study. *Physical & Occupational Therapy in Geriatrics* 21, 4 (2004), 1–19.

[40] Marijke Rensink, Marieke Schuurmans, Eline Lindeman, and Thora Hafsteinsdottir. 2009. Task-oriented training in rehabilitation after stroke: systematic review. *Journal of advanced nursing* 65, 4 (2009), 737–754.

[41] Brandon Rohrer, Susan Fasoli, Hermano Igo Krebs, Richard Hughes, Bruce Volpe, Walter R Frontera, Joel Stein, and Neville Hogan. 2002. Movement smoothness changes during stroke recovery. *Journal of Neuroscience* 22, 18 (2002), 8297–8304.

[42] 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.

[43] César Souza, Andrew Kirillov, Diego Catalano, and Accord.NET contributors. 2014. The Accord.NET Framework. https://doi.org/10.5281/zenodo.1029480

[44] Erik Stone and Marjorie Skubic. 2011. Evaluation of an inexpensive depth camera for in-home gait assessment. *Journal of Ambient Intelligence and Smart Environments* 3, 4 (2011), 349–361.

[45] Chuan-Jun Su, Chang-Yu Chiang, and Jing-Yan Huang. 2014. Kinect-enabled home-based rehabilitation system using Dynamic Time Warping and fuzzy logic. *Applied Soft Computing* 22 (2014), 652–666.

[46] 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).

[47] R PS Van Peppen, Gert Kwakkel, Sharon Wood-Dauphinee, H JM Hendriks, Ph J Van der Wees, and Joost Dekker. 2004. The impact of physical therapy on functional outcomes after stroke: what's the evidence? *Clinical rehabilitation* 18, 8 (2004), 833–862.

[48] Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 694–699.

[49] Liping Zhao, Xiong Lu, Xianglin Tao, and Xiaoli Chen. 2016. A Kinect-Based Virtual Rehabilitation System through Gesture Recognition. In *Virtual Reality and Visualization (ICVRV), 2016 International Conference on*. IEEE, 380–384.

[50] Wenbing Zhao, Hai Feng, Roanna Lun, Deborah D Espy, and M Ann Reinthal. 2014. A Kinect-based rehabilitation exercise monitoring and guidance system. In *Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on*. IEEE, 762–765.