A machine-learning model for automatic detection of movement compensations in stroke patients

Shir Kashi[1], Ronit Feingold-Polak[2], Boaz Lerner[3], Lior Rokach[1], Shelly Levy-Tzedek*[2,4,5]
CORRESPONDING AUTHOR: Shelly Levy-Tzedek (shelly@bgu.ac.il)

Abstract — During the process of rehabilitation after stroke, it is important that patients know how well they perform their exercise, so they can improve their performance in future repetitions. Standard clinical rating conducted by human observation is the prevailing way today to monitor motor recovery of the patient. Therefore, patients cannot know whether they are performing a movement properly while exercising by themselves. Adhering to the exercise regime makes the rehabilitation process more effective and efficient, and thus a system that can give the patients feedback on their performance is of great value. Here, we build a machine-learning-based automated model that give patients accurate information on the compensatory (undesirable) movements that they make. To construct the model, we recorded movements from 30 stroke patients, who each performed 18 movements, used to identify the presence of six types of compensatory movements in stroke patients' movement trajectories. We used the random-forest algorithm for training this multi-label classification model. We achieved 85% average precision across the six movement compensations. This is the first study to automatically identify movement compensations based on stroke patients' data. This model can be adapted for use in in-clinic and at-home exercise programs for patients after stroke.

Index Terms— Compensations, machine learning, multi-label classification, RAkEL algorithm, random forest, stroke rehabilitation, time series

1 Introduction

1.1 The importance and limitations of self-exercise for stroke patients

The intensity and repetition of post-stroke training are key to the efficacy of the rehabilitation process [1]. Up to 77% of stroke survivors experience upper limb (UL) impairment, which affects their function and reduces health-related quality of life [2]. For effective and efficient rehabilitation of their upper limb functionality, self-exercising in between physical therapy sessions is vital, and yet, many patients do not follow their exercise regime, which can hamper their recovery [2]. One explanation for why compliance rates are low is that patients undergoing rehabilitation are not able to assess their own functional state and their performance without the therapist [3, 4]. One of the major functional goals of rehabilitation after stroke is to retrain the coordination of reach-to-grasp (RTG) movements (e.g., in order to pick up a cup to drink from) [5]. In individuals with stroke, goal-directed movements are characterized by slowness, spatial and temporal discontinuity and abnormal patterns of muscle activation and joint synergy [6-8]. Individuals with stroke were reported to have less smooth [9], less accurate and less efficient RTG movements compared to healthy individuals [10], as was measured by the index of curvature [11] and by the jerk [12] of their movements. Following a stroke, patients who are not able to coordinate their muscle-activation patterns to perform an RTG task as they did before they had a stroke, develop compensatory movement patterns – e.g., bending their trunk, rather than extending their elbow – to reach an object located at arm's length. Several such compensatory characteristics of movement have been described in RTG tasks, both in the trajectory and in the interjoint coordination of the movement [6, 8].

1.2 From standard clinical rating to automated assessment

Cirstea and Levin (2000) demonstrated the importance of "knowledge of performance" in the upper limb rehabilitation process of stroke patients. That is, while the person is

- *S.K. is with the Department of Software and Information Systems Engineering at the Ben-Gurion University of the Negev. E-mail: shirik3@gmail.com*
- *R.F. is with Recanati School for Community Health Professions, Department of Physical Therapy, Ben-Gurion University of the Negev, Beer-Sheva, Israel. E-mail: polakr@post.bgu.ac.il*
- *B.L. is with the Department of Indusrtial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva, Israel. E-mail: boaz@bgu.ac.il*
- *L.R. is with the Department of Software and Information Systems Engineering at the Ben-Gurion University of the Negev. E-mail: liorrk@post.bgu.ac.il*
- *S.L. is with the Department of Physical Therapy, and the Zlotowski Center for Neuroscience at the Ben-Gurion University of the Negev, and with the Freiburg Institute for Advanced Studies (FRIAS) at the University of Freiburg. E-mail: shelly@bgu.ac.il*

practicing RTG movements, they need to know how well they performed the movement in order to improve their performance in future repetitions. The prevailing way to monitor motor recovery of the patient is carried out by direct human observation, using standard clinical ratings such as the Fugl-Meyer assessment (FMA), Functional Test for the Hemiplegic Upper Extremity (FTHUE), and the Brunnstrom stage [13, 14]. However, whereas those clinical scales are efficient tools, they also have some drawbacks [13-17]. The subjective judgement exercised by therapists using those tools may lack accurate quantifiable data [14], likely because it is difficult for the human eye to detect variations in a small-scale movement [17]. This makes it difficult to precisely evaluate gradual improvements in movement execution. In addition, the use of those tools is labor intensive and takes a considerable amount of time (at least 30 minutes) [2, 13, 15]. Furthermore, it is not suitable in the home settings as patients undergoing rehabilitation at home are not able to assess their own functional state with the FMA tool or other similar tools without a therapist [3, 4]. In order to help stroke patients to continue the rehabilitation training after they leave the hospital, there is a need for automated assessment [4]. Automated assessment holds several benefits in comparison to assessment by human observation only in clinics [14]. For example, it can offer a more detailed tracking of the time course of recovery by identifying variations in their movements pattern [14, 17], it avoids the "test" situation, which is often not representative of everyday function [14, 18], and also, it can make the assessment more objective by giving an automated score from a model instead of from a specific therapist [17]. In addition, the ability to make a quick and accurate evaluation could enable an efficient utilization of stroke-care resources, with clinician time being dedicated mainly to treatment, while the assessment is automated, even in the clinic [2, 15, 19].

1.3 Related work

To address this issue, a number of studies proposed a framework for automating upper limb assessments for stroke patients by using various sensors and classification schemes, most of them being based on machine-learning algorithms [2-4, 14, 16, 19-22]. The main purpose of most of these studies was to give an evaluation score per movement performed by the user. Otten et al. (2015) introduced an evaluation model using an artificial neural network (ANN) classifier, which outperformed the support vector machine (SVM) classifier. They used various sensors, such as GPS sensors, direction sensors (i.e., magnetic compasses), and acceleration sensors (i.e., accelerometers), from an Android-based smartphone in order to record movements of eight healthy participants. They asked the participants to perform all movements in three ways: faultlessly, partially, and not at all (motionless). From these, the authors calculated a set of movement features, such as elbow flexion, limb orientation, and joint angles. These features were used to determine a score for the participant's upper limb functionality, with a score of zero indicating the participant cannot perform any movement, and a score of two indicating they can perform the movement faultlessly. Kwapisz et al. (2011) also used an ANN classifier, but it was used to evaluate functional movement of the lower limbs, such as walking, going up or down stairs, jogging, sitting, standing, and not for movements of the upper limbs. That study also included only healthy participants [22]. To reduce the number of parameters required in the ANN model, Yu et al. (2016) used extreme learning machine (ELM)-based ensemble regression model and compared its results to the results of an SVM algorithm. They proposed to monitor the functional movement of the upper limb and attempted to predict the user's FMA score using sensor data. They found no obvious difference between the SVM and ELM algorithms in terms of accuracy. Importantly, they found that feature selection - i.e., narrowing down the feature space to the most informative set of movement features - leads to significantly improved accuracy [4].

There is limited information on automatically identifying the specific compensations in stroke patients' movements. Tormene et al. (2009) used dynamic time warping (DTW) and open-ended DTW (OE-DTW) to provide real-time feedback to neurological patients undergoing motor rehabilitation. They generated a dataset of multivariate time series from a sensorized long-sleeve shirt. One of the experiments they conducted was to recognize incorrectly performed movements, and in those, identify the specific error that was performed. However, there was only one healthy participant in this experiment, whom they asked to perform very slow movements or mimic two options of compensatory actions: adduction of the upper limb on a frontal plane or on a sagittal plane [23]. Similarly, Kizony et al. (2014) proposed a system that was designed to provide a home-based tele-rehabilitation program based on one healthy participant performing compensatory and non-compensatory movements.

Previous works show that ANN, SVM, ELM, and DTW could be beneficial for solving tasks such as multi-class

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2988945, IEEE Transactions on Emerging Topics in Computing

classification and regression [2-4, 14, 16, 19, 21, 22] . However, these algorithms are not suitable for more complicated tasks, such as multi-label classification tasks, when attempting to identify more than a single component of the movement (e.g., both excessive bending of the trunk and elevation of the shoulder). Furthermore, in order to build a model that identifies movement components that are found primarily in patients' movements, it is vital to collect the movement data from the relevant patient population, rather than from healthy individuals. However, most of the models built by previous works were based on movements of healthy participants and not on movements of post-stroke patients [21, 22]. Moreover, the number of the participants in the study is also a major factor in generating a representative model. Yet, previous works had between 1-7 participants, which may have limited their applicability [3, 21, 23, 24]. Finally, in order to build as accurate a model as possible, it is important to use high-precision sensors. Using a Kinect camera is rather common [2, 24], being a readily available and relatively affordable tool, though it is limited in its capacity to correctly detect fine motions [2], which then limits the overall accuracy of the generated model.

1.4 Our study

Here, we used data from 30 stroke patients, using a high-precision motion-capture system, to generate a multi-label classification model to detect movement compensations. To build such a model, it is necessary to choose the appropriate method for this multi-label task, which is more complex than those mentioned earlier, which had only a single outcome (for example, the presence of a single compensation vs. multiple concurrent compensations). According to Tsoumakas and Katakis (2007), there are two main categories of multi-label classification methods: transformation methods and algorithm-adaptation methods. The first category transforms the multi-label classification problem into one or more single-label classification problem(s), while the second adjusts known single-label classifiers to handle multi-label data [25, 26]. Since a main weakness of algorithm-adaptation methods is that they are mostly tailored to a specific classifier (e.g., SVM or decision tree), they lack the ability to be generalized, and thus the transformation methods perform better in this respect [25, 26].

To date, no algorithm has been developed to automatically identify the type of compensatory movements performed by actual patients who suffer from neurological conditions, such as stroke. Here, we propose such an algorithm, with which compensatory movements can be automatically detected without the need for an on-site clinician to be present. The machine-learning model we present here (1) will allow the patients to practice the desired exercise movements, as instructed by the therapist, and avoid performing undesirable movement patterns known as "bad learned use" [27] during self-practice, by providing accurate information on what specific compensations they performed (e.g., elevation of the shoulder); (2) will enable the therapist to receive information on the patient's at-home performance, in order to precisely adapt the overall training program to the patient's current ability; and (3) will serve as a personal ecological performance-assessment tool. In the future, the algorithm can be used to give recommendations for updating the exercise program during a session, in accordance with the patient's performance.

2 METHOD

2.1 Movement-compensation detection

The algorithm we present here will be used to identify the exact set of compensations the patient performed in each movement. Thus, we are dealing with a multi-label classification task, since any given movement can have between zero and six compensations in parallel. Figure 1 shows the process of generating the model, which we briefly overview here, and explain in detail below. To build the classification model for this problem, we first collected the movement data from stroke patients (Data Collection, Fig. 1.1). After that, data were analyzed in the Feature-Generation phase (Fig. 1.2). Movement features were generated using two methods in parallel and then combined: (1) biomechanics-inspired handcrafted features based on the motor-control literature, and (2) automatically extracted features by a dedicated software (the tsfresh package, see below). Then, Feature Selection (Fig. 1.3) was conducted in order to obtain an optimal set of features that will be the input to the random k-label sets (RAkEL) for multi-label classification algorithm (Fig. 1.4), which is suitable for multi-label problems. Finally, in the Evaluation phase (Fig. 1.5), we tested the performance of the model we built.

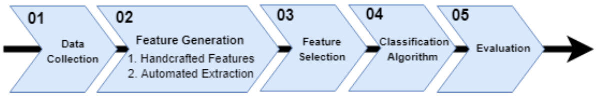


Figure 1. The movement-compensation detection process

2.2 Data Collection

2.2.1 Participants

Thirty post-stroke patients were recruited for this study from the "Beit Hadar" Rehabilitation Center (14 females, 16 males, mean age 70.3±9.4 years). The study was approved by the Barzilai Medical Center's Helsinki Committee, and all participants signed an informed consent form to participate in this study. Demographic and clinical information of the participants is listed in Table S1 (see Supplementary Materials).

2.2.2 Procedure

The participants were examined by a physical therapist between one to two weeks before their discharge date from the rehabilitation center. The examination was performed while the participants sat in front of a height-adjustable table. Participants were instructed to reach their impaired arm at a self-selected speed, forward, toward a cup located on the table, lift it, and place it on top of a 5 cm-high block, positioned on the table (Fig. 2). The participants were instructed to avoid bending their trunk as much as possible during the reach movement, but no restraint of the trunk was applied. RTG was performed at three different heights: (A) low - the height of the wrist when the hand is extended downwards, (B) medium, ~75 cm from the floor, the height of a standard table, and (C) high - the height of the participant's shoulder (Fig. 2). The cup was placed at an arm's distance, measured from the lateral acromion to the radial styloid process, to avoid excessive trunk movement during the reach movement. Reach and grasp movements were executed using an empty cup (273 gr) in half of the trials, and a cup filled with water (443 gr) in the other half (Fig. 2). Every combination of cup height and weight was repeated three times for a total of 18 RTG movements (3 heights x 2 weights x 3 repetitions). The order of the heights and weights was randomly set in order to prevent the influence of fatigue on particular combinations of height and weight.

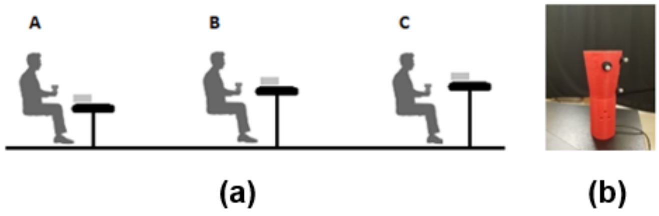


Figure 2. The experimental setup for the data-collection phase. (a): Participants were asked to reach to a cup placed on a table, pick it up, and place it back on a 5-cm block on top of the table. The table was set at three different heights: (A) low (~50 cm from the floor); (B) intermediate (~75 cm from the floor); and (C) high (~86-100 cm from the floor, depending on shoulder height). (b): The custom-built cup, embedded with a force sensor, with the three position markers (see text).

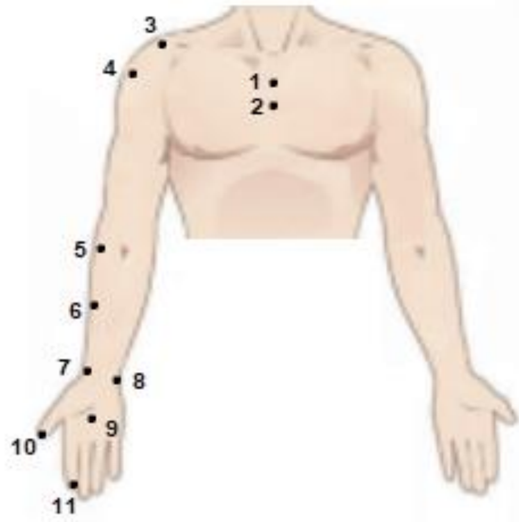


Figure 3. The location of the 11 body markers placed on each participant. (1-2: sternum, 3: shoulder, 4: proximal humerus, 5: elbow, 6: the middle forearm, 7-8: radial and ulnar styloid processes, 9: wrist, 10: thumb, and 11: index finger)

2.2.3 Motion-capture system

Position of the upper extremity joints during the RTG movement was recorded using a motion capture system V120: Trio (OptiTrack, NaturalPoint, Inc., OR, USA). The V120: Trio tracking system is a portable multiple-camera with a 6DoF optical object tracking technology. Eleven reflective markers were placed on the participants' upper body (Fig. 3). Markers were placed as follows: two markers were placed vertically aligned on the sternum to reflect the trunk motion (Fig. 3, Points 1-2), and one marker was placed on each of the following anatomical landmarks: lateral portion of the acromion [reflecting the scapular motion [28]] (Fig. 3, Point 3), proximal humerus (Fig. 3, Point 4), lateral epicondyle of the elbow (Fig. 3, Point 5), the middle forearm (Fig. 3, Point 6), radial and ulnar styloid processes (Fig. 3, Points 7-8), the dorsal side of the palm at the axis along the middle part of the third metacarpal bone (reflecting the wrist motion) (Fig. 3, Point 9), thumb (Fig. 3, Point 10), and index finger (Fig. 3, Point 11). Two additional markers were placed vertically on the wall behind the participant to serve as stationary reference points, and three additional markers were placed on the cup and defined by the system as a rigid body, so that the cup location can be tracked during each recording. Data sampling frequency of the Trio system is 120 Hz.

2.2.4 Motion-capture system

Grip forces were measured with a 3D force sensor (Nano25-E Transducer, ATI Industrial Automation, INC) embedded in a custom-built 3D-printed cup (Fig. 2). The data sampling frequency of the force sensor was 100 Hz.

The output from the force sensor was the summed grip forces applied on the cup.

2.3 Feature Generation

To build an initial set of features from the collected data, we used two different approaches. The first is based on the analysis of the reach-grasp-lift movement by calculating (with Matlab R2017b) first the segments of each movement and then a wide variety of biomechanical hand-crafted metrics, including velocity, jerk, index of curvature, angels of the joints, etc. derived from the time series data generated using the markers and the force values measured during the movement. The movement data were down-sampled to 100 Hz to match the sampling frequency of the force data. The second approach is based on time-series feature extraction using scalable hypothesis tests implemented by the tsfresh package [29].

2.3.1 Movement segmentation

RTG is divided into a transport component, which is the change in position of the hand over time, and a grasp component, which is the change of the distance between the index finger and thumb over [30]. In healthy individuals, certain elements in reaching and grasping display invariant behaviors suggesting key principles in motor control. For example, movement trajectories involving more than one joint tend to be straight, smooth, and have bell-shaped velocity profiles [31]; Peak deceleration point usually occurs around the time of object contact [30]; The start time of the opening of the hand is correlated with the start time of hand movement toward the object, and the time of maximum hand opening is correlated with the time of peak deceleration of the hand [30]. Smoothness of the movement is widely regarded as a hallmark of coordinated movement. Jerk, the third derivative of position with respect to time, has been used as an empirical measure of this quality [32]. We automatically segmented each of the participants' movements into three segments, as shown in Figure 4, based on four time points (T1-T4), according to the phase of the movement: reaching to the cup (REACH), grasping the cup (GRASP), and raising the cup and placing it on the block (LIFT):

1. T1 – start of movement – the time at which 10% of the maximal velocity of the wrist is reached [6].
2. T2 – start of grasp – the time at which the participant applied force equal to 5% of the difference between the maximal and minimal forces applied on the cup in a given trial.
3. T3 – start of lift – the time at which the cup reached 10% of its maximal height during the trial.
4. T4 – end of movement – the time at which the movement ended, calculated as follows: the force trace was scanned from the end of the trial backward until the first time the value of the force was 20% of the maximal force during that trial. Then, the force trace was scanned from this point forward, to find the first time the value of the force was 5% of the maximal force during that trial. That point in the time series was set to be T4.

In each movement, T2 was first identified, then T1 was calculated in the interval between the start of the recording and T2, followed by T3, and then T4. Once these values were determined, the segmentation was applied to the entire time series of the movement, across all markers. The segmentation rules we detail above were informed by previous literature [33], and adapted, where needed, to the movement and force profiles generated when the stroke patients perfomed the RTG movements to different heights with different weights. We verified that the time points T1-T4 correspond to the appropriate movement phases, as shown in Fig. 5.

2.3.2 Biomechanical handcrafted movement features

Kinematic features of the movement were calculated from six main categories: ‘Jerk’, ‘Velocity’, ‘Angles’, ‘Aperture’, ‘Curvature’, and ‘Force’, all of which are known to have different characteristics in movements of individuals after stroke [6, 7, 9, 10, 12, 30, 34]. Table 1 lists the features we calculated from each category, the markers used to derive each feature, and for which segments of the movement the features were calculated (see Section 2.3.1): REACH (T1 to T2), GRASP (T2 to T3), LIFT (T3 to T4), and ALL (a unified segment that begins in T1 and ends in T4). The combination of two segments indicates a segment that is composed of both (e.g., GRASP + LIFT between T2 and T4).

In the ‘Angles’ category, six angles were calculated based on position data from the relevant markers:

1. Scapula elevation (ScapulaEle) – the angle between the two sternum markers and the shoulder marker (Fig. 3, Points 1-3).
2. Scapula rotation (ScapulaRot) – the angle between the bottom sternum marker and the shoulder marker (Fig. 3, Points 2-3).
3. Trunk– the angle between the two sternum markers (Fig. 3, Points 1-2) and the wall markers.
4. Elbow–the angle between the proximal humerus, elbow, and middle forearm markers (Fig. 3, Points 4-6).
5. Shoulder– the angle between the upper sternum, shoulder, and proximal humerus markers (Fig. 3, Points 1, 3-4).
6. Wrist– the angle between the middle forearm, the radial styloid process, and the wrist markers (Fig. 3, Points 6,7,9).

The correlations between pairs of angles were then calculated per movement, as well as between the angles and the aperture of the hand (see Table 1 for the list of pairs for which correlations were calculated).

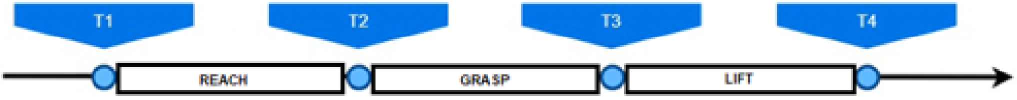
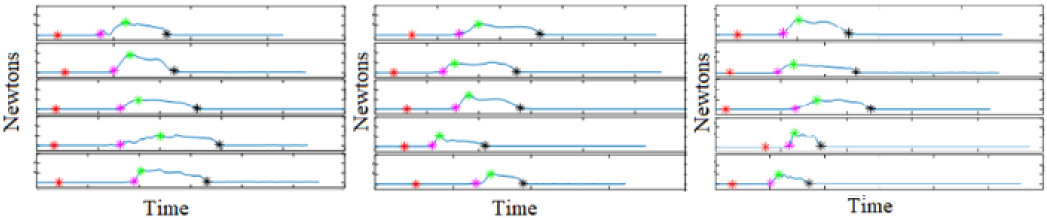


Figure 4. The process of movement segmentation



F igure 5. Force profiles during a single trial from three different movements (columns) performed by five different participants (rows). The red, magenta, green, and black asterisks represent T1, T2, T3, and T4, respectively.

TABLE 1
The handcrafted features

Category	Feature	Markers/Sensors	Segments
Jerk	**Mean Squared Jerk (MSJ) –** $\frac{1}{\vert T\vert}\int \frac{1}{2}(\mathbf{jerk}^2)\ /\ \mathbf{max}(\mathbf{velocity})$ *Velocity* - the first derivative of position with respect to time for the duration of the movement, T. *Jerk* -the third derivative of position with respect to time for the duration of the movement, T.	Wrist	REACH, GRASP, LIFT, ALL
Velocity	**Average, Maximum, Minimum, Time to Maximum**	Wrist, Elbow	REACH, GRASP, LIFT, ALL
Angles ScapulaEle ScapulaRot Trunk Elbow Shoulder Wrist	**Correlations –** Trunk ↔ Elbow Trunk ↔ Aperture Trunk ↔ Shoulder Trunk ↔ ScapulaEle Trunk ↔ ScapulaRot Trunk ↔ Wrist ScapulaEle ↔ Elbow ScapulaEle ↔ Aperture ScapulaRot ↔ Aperture Wrist ↔ ScapulaEle Wrist ↔ Elbow	Sternum 1-2, Shoulder, Humerus, Elbow, Forearm, Radial, Wrist	REACH, GRASP, LIFT, ALL
	Maximum, Minimum, Average	Trunk, Elbow, Shoulder, Wrist, Sternum 1-2	REACH, GRASP, LIFT, ALL
Aperture	**Maximum Distance –** between the positions of the thumb and index-finger markers.	Thumb, Index	REACH
	Time to Maximum Distance – The time elapsed from T1 to the time at which the participant reached the Maximum Distance.	Thumb, Index	REACH
	Time to T2 – The time elapsed from when the participant reached the Maximum Distance to T2.	Thumb, Index	REACH
	Std1– Standard deviation of the aperture amplitude between T1 and the time that 'Maximum Distance' was reached.	Thumb, Index	REACH
	Std2 – Standard deviation of the aperture amplitude between the time that 'Maximum Distance' was reached and T2.	Thumb, Index	REACH
Curvature	**Straight-line distance –** The length of a straight line connecting the locations of the wrist marker at T1 and T4.	Wrist	REACH, GRASP+LIFT
	Path length – The actual path length of the movement made between each segment.	Wrist	REACH, GRASP+LIFT
	IC – Index of curvature: the ratio between the Straight-line distance and Path length.	Wrist	REACH, GRASP+LIFT
Force	**Summed forces – Measured** across all durations of movement.	Force sensor	GRASP+LIFT
	Time to max force – Time duration between T2 and the time at which the maximum force was reached.	Force sensor	GRASP, LIFT
	Segment duration – Duration of a movement segment.	Force sensor	GRASP, LIFT, GRASP+LIFT
	Duration ratio – The ratio between the segment durations of GRASP and LIFT.	Force sensor	GRASP, LIFT
	Variance – Variance of the force measured during the segment.	Force sensor	GRASP, LIFT
	Time duration max aperture – The time duration between the time max aperture was reached and T2, divided by the variance in segment.	Force sensor	GRASP, LIFT
	Average force	Force sensor	GRASP, LIFT, GRASP+LIFT
	Std – standard deviation of the difference between any two consecutive force values.	Force sensor	GRASP+LIFT
	Variance X **Segment duration**	Force sensor	GRASP, LIFT
	Average X **Segment duration**	Force sensor	GRASP, LIFT

The first, second, third, and fourth columns list the main category the feature belongs to, its meaning, the markers/sensors that were used for its calculation, and which segments the calculation was applied to, respectively.

2.3.3 Automated extraction of time-series features

We used the Python 3.6 tsfresh package to generate 3,000 time-series features automatically [29]. This package filters the features with respect to their significance for the classification task, while controlling the expected percentage of selected but irrelevant features. The features it extracts describe basic characteristics of the time series such as the number of peaks, average or maximal value of a signal or more complex features such as the time reversal symmetry statistic. The features belong to one of three main categories [29].

1. Summary statistics- such as: maximum, minimum, mean, variance, standard deviation, skewness, kurtosis, length, median, quantile of empiric distribution.
2. Sample distribution - such as: absolute energy, augmented Dickey-Fuller test statistic, binned entropy, distribution characteristics, symmetry, mass quantile, number of data points above mean/median, and number of data points below median.
3. Observed dynamics – such as: autoregressive integrated moving average (ARIMA) model coefficients, continuous wavelet transformation coefficients, and fast Fourier components [29].

For the sake of brevity, we elaborate here on six of these features, which we found to be most influential for the performance of the model:

1. Autocorrelation – An autocorrelation estimator is given in Eq. (1), where n is the length of the time series x_i, σ^2 its variance and μ its mean. l denotes the lag between observations:

$$\frac{1}{(n - l)\sigma^2}\sum_{t=1}^{n-l}(x_t - \mu)(x_{t+l} - \mu) \tag{1}$$

2. Change quantiles mean – The mean absolute value of consecutive changes of the series x inside a range defined by the upper and lower quantiles of the distribution of x.
3. Change quantiles variance – The variance absolute value of consecutive changes of the time series, excluding extreme values (defined by a quantile range).
4. Energy ratio by segments – The sum of squares of segment i out of N segments expressed as a ratio with the sum of squares over the whole series. N is the number of segments to divide the series into and i is the segment number (starting at zero) to return a feature on.
5. Abs energy – The absolute energy of the time series which is the sum over the squared values of the examined time series.
6. Larger standard deviation – Boolean variable denoting if the standard deviation of x is higher than r times its range, which is the difference between maximum and minimum of x.
7. Ratio beyond $r\sigma$ – Ratio of values that are more than $r * std(x)$ (so they are $r\sigma$) away from the mean of x divided by the length of the time series.

The time-series sequence we used as input to the tsfresh package is the traces of the three position axes (x, y, z) of the wrist marker (Fig. 3, Point 9) for each participant for each of the 18 RTG movements. The wrist marker was chosen as it is common to use this marker for kinematic data analysis [34, 35].

2.3.4 Compensation labeling

Two expert physical therapists labeled each of the 18 movements per participant with the set of compensations that participant made, if any. To label the movements, they viewed a visualization of the collected data (videos of the movement created by the markers). Compensations were only labeled when both physical therapists agreed they were present, following a joint discussion. The possible compensations were: trunk-flexion, scapula-elevation, scapula-rotation, shoulder-flexion, elbow-flexion, distal dys-synergy.

2.4 Feature Selection

It was necessary to perform feature selection on the large set of features we generated using both methods (over 3,000) for two reasons: (1) to avoid over-fitting due to a large number of features compared to the number of training instances (movements); and (2) to identify the most meaningful features for a leaner and more efficient model. Feature selection allowed us to reduce the dimensionality of the feature space, and remove redundant, irrelevant, and noisy features, to enable the classification algorithm to be more accurate and rapid. As will be described in Section 2.5, we classified movements by multi-label classification, which transforms the problem into one or more single-label classification tasks [36]. We used the WEKA 3.8 program [37] for this task. WEKA is a workbench for machine learning that is intended to aid in the application of machine-learning techniques, as the feature selection phase [37]. We applied, for each single label (i.e., compensation), feature selection that searches the space of feature subsets by greedy hill climbing augmented with a backtracking facility. The features were evaluated by 'CFS' (correlation-based feature selection) [37], which evaluates the contribution of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Therefore, subsets of features that are highly correlated with the label (compensation) while having low crosscorrelation among them are preferred [37].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.2988945, IEEE Transactions on Emerging Topics in Computing

We applied this selection approach on each of the six compensation labels separately (section 2.3.4) and kept features that were selected for between two and five labels, since there were too many features that appeared only in a single label set. None of the features were selected for all the six labels. All the features we obtained in Sections 2.3.2 and 2.3.3 were combined to one set of 156 features.

2.5 The multi-label classification using RAndom k-labELsets (RAkEL) algorithm

Recall that a multilabel classification algorithm is necessary when attempting to identify more than a single compensation in the movement. RAkEL is a popular ensemble approach for multi-label classification over a set of labels L (the six compensation types). To consider label correlations, it constructs an ensemble of label powerset (LP) classifiers, each considering a different subset of labels as a single label. Following this approach, the LP classifier learns one single-label classifier [38]. Since LP classifiers may have the disadvantages of having many different possible label sets, ($2^{|L|}$), each with only a small number of examples, each LP is trained using a different small random subset (k) of labels. For training, this approach uses the ensemble method - multiple learning classifires (M), to obtain better predictive performance [38]. The higher the values of M and k are, the better is the performance of the RAkEL method. However, since the complexity of RAkEL grows exponentially with the size of label sets k, but only linearly with the number of classifiers, it is common to set k to a small value. Note that if k=1, RAkEL trains M=|L| (=6) binary classifiers. We trained the RAkEL model using a *random forest* as the base classifier; the size of the label subset was k=4, and the number of classifiers was M =16 (see explanation below in the Results section). For the implementation, we used R-Studio version 3.5.1.

3 Evaluation

3.1 Leave-one-out cross validation

We used the leave-one-out cross validation (LOOCV), which is necessary when there is insufficient amount of data [39]. In LOOCV, each of the 30 models is trained on a different subset of 29 participants, and the data from the remaining participant was used as a validation set [39]. LOOCV is useful for avoiding the statistical problem of overfitting in models in which the same samples are used both for training and prediction, or when the number of instances is small and there are many variables [40]. Another property of the LOOCV is that it provides an almost unbiased estimate of the generalization ability of a classifier [39].

3.2 Evaluation metrics

Multi-label classification requires different metrics for evaluation than those used in traditional single-label classification [41, 42]. We used here the precision metric and the Hamming loss, which are widely used in the literature to evaluate multi-label classification methods [41, 42].

3.2.1 Micro and macro averaged precision

Micro and macro averaged precisions are two ways to calculate the precision over all the labels in a multi-label classification problem [43]. The precision can be computed globally over all labels, which is the "micro-averaged precision", or for each label separately and then be averaged over them, which is called "macro-averaged precision". Equations 2 and 3 were used to calculate the micro- and macro-averaged precision, respectively, while |L| represents the number of labels (i.e., compensations), TP represents the True Positive variable (i.e., the portion of correctly identified compensations), and FP represents the False Positive variable (i.e., the portion of incorrectly identified compensations) [43]. The range for both TP and FP is [0,1] while the optimal value for TP is 1 and the optimal value for FP is 0.

$$\textbf{Micro}-\textbf{averaged precision} = \frac{1}{|L|} * \frac{\sum_{i=1}^{|L|} TP_i}{\sum_{i=1}^{|L|} TP_i + \sum_{i=1}^{|L|} FP_i} \quad (2)$$

$$\textbf{Macro}-\textbf{averaged precision} = \frac{1}{|L|} * \sum_{i=1}^{|L|} \frac{TP_i}{TP_i + FP_i} \quad (3)$$

While micro-averaging can be used to know how the system performs overall across the data, macro-averaging is preferable if there is a class imbalance [5]. Providing both scores is more informative than providing either of them alone [43].

3.2.2 Macro-Averaged Hamming loss

The Hamming loss evaluates how many times an instance-label pair is misclassified; in other words, a label not belonging to the instance is predicted or a label belonging to the instance is not predicted [41]. The Hamming loss was computed as follows: for each movement, the predicted list of compensations was compared against the actual presence of each of the six possible compensations for that movement. Then, the number of mismatched pairs (of predicted compensation vs. actual presence) across all examined movements was divided by the number of possible compensations (|L| = 6) and by the number of movements examined (N). Table 2 shows an example of this computation.

$$\textbf{Macro}-\textbf{averaged Hamming loss} = \frac{1}{N} * \frac{\sum_{j}^{N} \sum_{i}^{L} [y_{pred}^{i} \neq y_{true}^{j}]}{|L|} \quad (4)$$

TABLE 2
Computation example of the Hamming loss metric

	i=1	i=2	i=3	i=4	i=5	i=6
y^i_{pred}	1	0	1	1	0	0
y^i_{true}	0	0	1	1	1	0
$y^i_{pred} \neq y^j_{true}$	1	0	0	0	1	0
$\frac{\sum_i^L [y^i_{pred} \neq y^j_{true}]}{\lvert L \rvert} = \frac{2}{6}$						

4 RESULTS

In the movement-compensation detection process, we use 156 features in total (both from handcrafted and tsfresh features), the result of the feature-selection phase (Fig 1.3). **Table 3** presents the results obtained from the model in three feature settings: using handcrafted features (45 features), tsfresh features (111 features) and both (all 156 features). As we can see from the table, macro and micro averaged precision were higher when all 156 features were used, with a macro-averaged precision of 0.85, and a false-positive rate of 0.15. Not only that, but also the Hamming loss was the lowest in that condition. The calculation of the macro-averaged precision metric as we explain in Section 3.2.1 is done by computing the precision for each label, i.e, compensation, separately before averaging them. **Figure 6** shows the precision for each label.

TABLE 3
Average predictive performance for different feature sets

Formula Features	Micro-averaged precision	Macro-averaged precision	Macro averaged Hamming loss
Handcrafted	0.78	0.81	0.21 $**^a$
Tsfresh	0.76	0.81	0.22 $*^a$
Handcrafted + tsfresh	0.81	0.85	0.19

[a] *The symbols **/* indicate that the difference between Handcraft + tsfresh is significantly better than the corresponding feature set at p<0.05/ p<0.1, respectively*

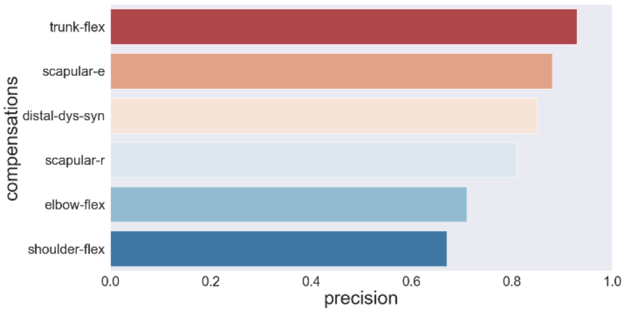


Figure 6. Precision scores for each of the six compensations: trunk flexion (trunk-flex, 93%), scapular-elevation (scapular-e, 88%), distal-dys-synergy (distal-dys-syn, 85%), scapular rotation (scapular-r, 81%), elbow flexion (elbow-flex, 71%) and shoulder flexion (shoulder-flex, 67%).

The best results obtained by the RAkEL algorithm were for k=4 and k=2. Since both give a micro-averaged precision of 0.81, but k=2 has a significantly longer running time, we opted for using k=4. **Table 4** shows a grid search of parameter k and the corresponding micro-averaged precision. It is common to determine the number of classifiers (M) of the RAkEL algorithm as the number of labels or twice this number (i.e., in this case, between L=6 and 2*L=12). Since the larger the number of classifiers, the more reliable is the result, we tested, using trial and error, values of M>12 to find the optimal value for M, and concluded by choosing M=16.

TABLE 4
Micro averaged precision for the k parameter in the range [1,6] in the RAkEL algorithm

	k=1	k=2	k=3	k=4	k=5	k=6
Micro-averaged precision	0.80	0.81	0.77	0.81	0.75	0.75

To test significance of the results in **Table 3**, we compared the macro-averaged Hamming loss, which is calculated for each movement of each patient, among the three features settings. First, we applied adjusted Friedman test [44] and rejected the null hypothesis that all settings perform the same (p-value<0.05). Once the null hypothesis was rejected, we used the post-hoc Nemenyi test in order to compare the settings with each other. The difference between the combined feature set and the handcrafted feature set was found to be statistically significant with p-value<0.05 (the combined is better). The difference between the combined feature set and the tsfresh feature set had a p-value<0.1. Finally, there was no statistically significant difference between the tsfresh feature set and the handcrafted feature set.

Figure 7 shows all 156 paired feature correlations. The matrix has the same number of rows and columns as the number of features. Cell (i,j) represents the correlation between features i and j. The small number of feature pairs with high correlation demonstrates that the features mostly contribute to the model individually, and not in combination with other features.

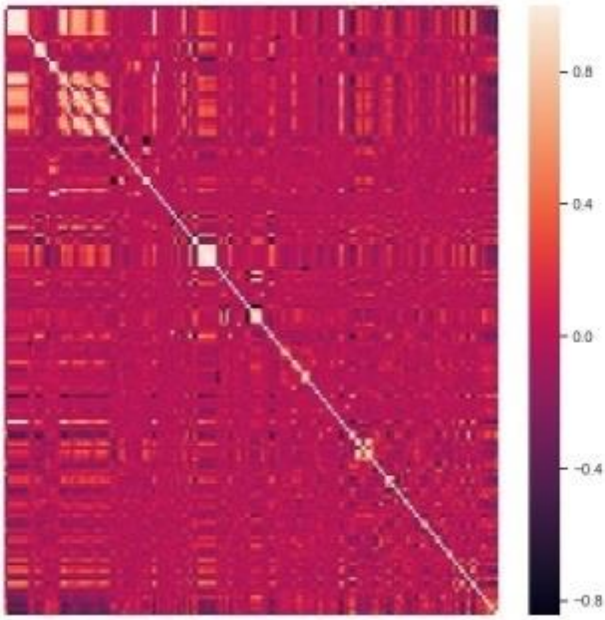


Figure 7. Pairwise correlations between all features. There are 156 rows and columns in this matrix, corresponding to the number of features used. Cell (i,j) represents the correlation value between features i and j.

To find which of the 156 features were the most important, we applied the LOOCV protocol to calculate the macro-averaged precision of the model, each time with a different feature absent. Then, we extracted the 10 features that have the greatest impact on the model – i.e., those without which, the model's macro-averaged precision score was lowest. **Table 5** lists the 10 most informative features, with three handcrafted features and seven which were generated by the tsfresh package. The tsfresh features were generated by the default hyper-parameter of the package without any optimization. **Figure 8** shows values of the macro-averaged precision score of the model without each of the top 10 features.

TABLE 5
The 10 most-important features

	Feature	**Approach**	**Score**
1	Minimum elbow velocity	Handcrafted	0.8268
2	X wrist autocorrelation	tsfresh	0.8268
3	Correlation wrist ↔ trunk	Handcrafted	0.8263
4	X wrist variance - change quantiles	tsfresh	0.8261
5	X wrist energy ratio by segments (10,1)	tsfresh	0.8258
6	X wrist abs energy	tsfresh	0.8255
7	Mean force	Handcrafted	0.8252
8	Y wrist large std	tsfresh	0.8251
9	X wrist energy ratio by segments (10,6)	tsfresh	0.8249
10	Ratio beyond r sigma	tsfresh	0.8247

The description, source (handcrafted or based on tsfresh), and average precision score in the absence of each of these features

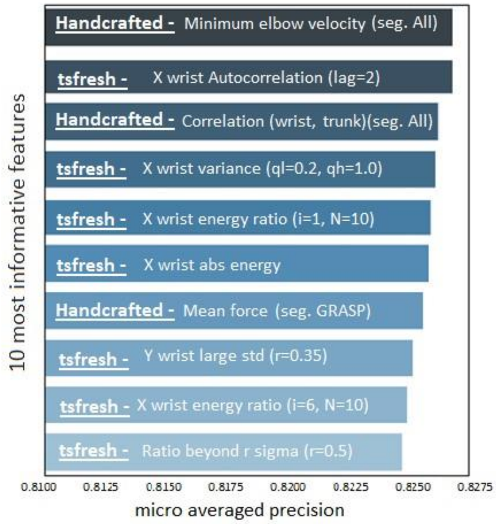


Figure 8. The micro-averaged precision of the 10 most informative features

Table 6 lists the top-5 features for each table height. **Table 7** lists the top-5 features for each compensation.

TABLE 6
The 5 most important features per table height (low/medium/high)

	Low	**Medium**	**High**
1	Minimum elbow velocity (H)	X wrist variance – change quantiles (T)	Minimum elbow velocity (H)
2	X wrist autocorrelation (T)	Minimum elbow velocity (H)	Max angle of scapulae (H)
3	Z wrist mean – change quantiles (T)	Y wrist large std (T)	X wrist variance – change quantiles (T)
4	Correlation wrist, trunk (H)	X wrist mean – change quantiles (T)	Correlation trunk, scapulaRot (H)
5	X wrist energy ratio by chunks (T)	Ratio beyond r sigma (T)	Correlation wrist, trunk (H)

For each feature, it is noted (in parentheses) whether it was handcrafted (H) or generated by the tsfresh package (T).

TABLE 7
The 5 most important features per compensation

	C1: trunk flexion	**C2: scapular elevation**	**C3: scapular rotation**	**C4: shoulder flexion**	**C5: elbow flexion**	**C6: dys-synergy distal**
1	Minimum velocity, GRASP (H)	Correlation (trunk, elbow) (H)	Minimum elbow velocity, LIFT (H)	X wrist energy ratio by segments (T)	Time to maximum velocity, REACH (H)	X_wrist__autocorrelation (T)
2	Minimum elbow velocity, ALL (H)	Correlation (trunk, scapularRot) (H)	Std diff (H)	X_wrist-autocorrelation (T)	Minimum velocity, LIFT (H)	Time to maximum velocity, LIFT (H)
3	Correlation (trunk, shoulder) (H)	Correlation (trunk, wrist) (H)	Time to T2 (H)	Correlation (trunk, wrist) (H)	Minimum elbow velocity, ALL (H)	Std_diff (H)
4	Y_wrist__percentage_of_reoccurring_datapoints (T)	Correlation (wrist, elbow) (H)	X_wrist__agg_"var" segment (T)	Correlation (trunk, shoulder) (H)	Mean force, LIFT (H)	wrist_abs_energy (T)
5	Correlation (trunk, wrist) (H)	Mean force, LIFT (H)	X_wrist__augmented _"pvalue" (T)	Mean force, LIFT, (H)	Time to T2 (H)	Mean force, LIFT phase (H)

The most informative five features (rows) for each of the six compensations (columns) are listed in order from the most informative (top) to least (bottom). Each feature name is followed by an indication of whether the feature was handcrafted (H) or generated by tsfresh (T).

As an example of the information contribution of an individual feature, we show in **Figure 9** the value of the 'X wrist abs energy' feature for movements with and without each of the six compensations. As shown in the figure, movements with compensation tend to have lower values for this feature. One possible explanation for this is that, for those movements that included compensations, the compensation bypassed the need to extend the elbow, which lead less curvature in the motion of the wrist, resulting in a decrease in the energy of the wrist movement (for the energy calculation see the Methods section above)

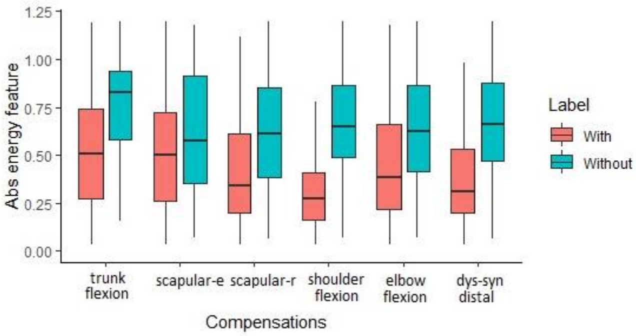


Figure 9. The 'abs energy' feature, generated by tsfresh, in movements with and without each of the six compensations. In red are the values of this feature for all the movements that include the compensation; In green are the values of this feature for all the movements without this compensation.

5 DISCUSSION

The main objective of this study was to construct an automatic algorithm that can identify whether and which compensations a post-stroke individual made when reaching to a cup, grasping, and lifting it. Our motivation in developing this algorithm was to enable individuals after stroke to practice everyday actions on their own, in addition to the practice they do during physical and occupational therapy sessions, by providing them with the information on whether they performed an undesirable compensation movement during their practice. In this experiment, we used data we recorded from position markers placed along the upper limb of 30 post-stroke individuals when they reached and grasped a cup placed at three different heights, with the cup being either empty or full. Feature generation and feature selection were made to find the optimal set of features which will result in the best performance of the algorithm. We produced two main sets of features: handcrafted and tsfresh-based features. Since the algorithm was trained to identify several compensations simultaneously made in an individual's movement, this is a multi-label task. Therefore, we chose the RAkEL algorithm, and achieved high classification rates, with 0.85 macro-averaged precision. That is, we identified the presence of the six main compensatory movements (described in detail in the Results section) in 85% of the cases, on average. The identification rate varies per compensation, and ranges between 67% and 93%. We found that the combination of the handcrafted features and the tsfresh features resulted in significantly more accurate identification rates, compared to using each set of features separately. The analysis we present in **Tables 5** and **6** provides interesting insights into the most informative features in identifying the six compensations. Although there is an overlap in the most informative features across table heights (**Table 6**) and compensations (**Table 7**), they are not exactly the same features in the same order of importance for each of the table heights or compensations. That is, there are some main features that will be informative for all table heights and for all six compensations (**Table 5**), as well as unique features that contribute specifically to the identification of a particular compensation, or to the idenfication of compensations performed at a certain table height.

The prevalent way to evaluate movements made by post-stroke individuals is by a physical therapist using standard clinical rating scales such as the FMA or the FTHUE [13]. Recently, an effort has been made to automate these clinical ratings using machine-learning algorithms [2-4, 16, 19-22]. The motivation for automating the ratings is threefold: (1) to save time – evaluation with a model can take less than a few minutes, whereas performing the clinical test may take 30 minutes [2, 13, 15]; (2) to improve the precision of the evaluation, as the human eye may not detect small-scale movements that can be detected by accurate sensors [17]; and (3) to help post-stroke individuals to receive an evaluation of their movement quality when a clinician is not available to provide one, e.g., during home practice [4]. Using this system for home practice can also help patients who refrain from going to the clinic due to low availability and accessibility, lack of knowledge of opportunities, high costs of organized activity, inclement weather, or who feel uncomfortable exercising in public since they are concerned about how others might perceive them [45]. The models that have been developed thus far are useful in providing a clinical score. In contrast, our model is based on performing functional tasks. Rather than generating a score, it provides information on the exact compensation the individual performed. We anticipate this will be highly pertinent and informative output for the patients' rehabilitation process, as they work on recovering the ability to perform everyday tasks.We found that when using either the handcrafted features or the tsfresh-generated features, there was no significant difference in the macro-averaged Hamming loss, indicating that both sets of features are equivalent in their contribution to identifying the presence of compensations. However, their combination – handcrafted with tsfresh features – significantly improved the performance of the model, compared to using each set separately. A possible reason for this is that when we created the RGL segments, which were chosen based on the motor-control literature, we attempted to automatically segment the movements across all participants. Since each participant performs movements differently, there was a trade-off between the accuracy of the segmentation, which affects the model's output, and automaticity of the data segmentation. Since automation is a key feature in our system, we strove for automatic segmentation, which may have resulted in the loss of some participant-specific information. The tsfresh set of features, which were calculated for each movement as a whole without applying our segmentation rules, apparently added information that was lost in the calculation of the handcrafted set. Thus, the combination of both sets of features led to the best results.

The algorithm we developed offers several benefits over existing models for movement evaluation: (1) It uses everyday functional movements, of the type individuals after stroke are often asked to practice; Thus, it avoids the artificial nature of a test situation which does not reflect real everyday movements [14, 17, 18]; (2) It is based on data from stroke patients; (3) It provides direct information on the compensatory movements that the individual performed – whether there were any, and which. Thus, it can address a potential concern that patients may have regarding at-home practice. According to Moriss et al. (2017), translation of motivation into actual activity depends on capability for physical activity. For example, intrinsic influences, such as the fear of negative consequences of physical activity, may prevent individuals from being physically active [45]. Indeed, repeatedly performing compensatory movements is known as "bad learned use" [27], and should be avoided. Using a model as the one we present here, provides the users with accurate feedback on their movement performance. It can assist patients to perform movements more correctly and thus help them experience success, which, according to Moriss et al (2017), leads to increased motivation and is translated to confidence in the general capability to be active.

By tracking the individuals' performance over time (e.g., what compensations are present, and whether they diminish over time), both the clinical team and the patient can have an accurate evaluation of the process of recovery and identify particular recurring difficulties.

5.1 Limitations

One possible limitation of our work is the sample size – 30 post-stroke individuals. While it is large compared to previous works with human participants, a larger sample size may allow for better generalization, and have a more balanced data set (a better representation of the different impairment levels) [46]. In addition, this group of participants was heterogeneous in terms of their functional ability, with patients displaying low, moderate and severe levels of impaired function. Despite this, the model we developed reached a macro-averaged precision of 0.85.

Another limitation is the use of markers for data collection. In this work, we aimed at generating an accurate algorithm for the identification of compensatory movements in stroke-patient movements. For that purpose, we used high-precision motion-capture device (V120: Trio OptiTrack, NaturalPoint, Inc., OR, USA, accuracy ≤ 1.0mm). It is conceivable that simpler implementations of the model may use a smaller number of sensors (in our model, we used only 6 of the 11 sensors we recorded from, namely: the sternum 1, sternum 2, shoulder, elbow, wrist and force sensors), or potentially a different, low-cost, simple sensor system, which would add to the user's convenience of using the algorithm both in terms of procedure and of price. Enabling a low-cost automated supervision of at-home practice would be important in that during home-based practice there are fewer constraints on time and space, so the patients can practice more frequently, for longer periods of time, and also according to their own schedule and terms [3, 4, 47]. Also, at home, the individuals can engage in more ecological exercises that are compatible with their everyday routines, which may be more indicative and useful from an evaluation session in the clinic [3, 4, 47].

6 Summary and Future Directions

In this study, we constructed a model that identifies the presence of compensations in stroke patients' movements, to be used in the process of rehabilitation. We achieved 85% macro-averaged precision across the six movement compensations we studied. This is the first study to identify compensations based on stroke patients' data. Here, we used a high-precision movement-capture system. However, future work with a more affordable sensor system may open the possibility for stroke patients to use the model system for home-based training. Such an affordable and simple-to-use tracking system, which can provide real-time position information would be necessary for an in-clinic or at-home implementation. The automated algorithm we present here may further be combined with socially assistive robotics (SAR), which can administer the exercise set, and provide feedback on the user's performance [48-51]. A potential line of future investigation would be how the compensation-specific information that the SAR may deliver to users affects their acceptance and level of trust in using such a device in the process of rehabilitation [52, 53]. Lastly, it would be instructive to collect and use data from the unimpaired arm when individuals with stroke perform functional RTG movements, as compensatory strategies may also involve the unimpaired side of the body.

Acknowledgment

The authors would like to thank Anna Yelkin for collecting the patient data and labeling the compensations. The research was supported in part by the Helmsley Charitable Trust through the Agricultural, Biological and Cognitive Robotics Initiative and by the Marcus Endowment Fund, both at the Ben-Gurion University of the Negev. Financial support was provided by the Rosetrees Trust, the Consolidated Anti-Aging Foundation and the Borten Family Foundation grants. This research was also supported by the Israel Science Foundation (grants No. 535/16 and 2166/16) and the Israel National Insurance Institute and received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754340.

References

[1] J. Brackenridge, L. Brandam, S. Lennon, J. Costi, and D. Hobbs, "A Review of Rehabilitation Devices to Promote Upper Limb Function Following

Stroke.," Neuroscience and Biomedical Engineering, vol. 4, pp. 25-42, 2016.

[2] S. Lee, Y. S. Lee, and J. Kim, "Automated Evaluation of Upper-Limb Motor Function Impairment Using Fugl-Meyer Assessment. ," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, pp. 125-134., 2018.

[3] J. Wang, L. Yu, J. Wang, L. Guo, X. Gu, and Q. Fang, "Automated Fugl-Meyer assessment using SVR model. ," presented at the IEEE International Symposium on Bioelectronics and Bioinformatics (ISBB), 2014.

[4] L. Yu, D. Xiong, L. Guo, and J. Wang, "A remote quantitative Fugl-Meyer assessment framework for stroke patients based on wearable sensor networks. ," Computer methods and programs in biomedicine, vol. 128, pp. 100-110, 2016.

[5] V. Van Asch, "Macro-and micro-averaged evaluation measures [[basic draft]]. ," presented at the CLiPS, Belgium, 2013.

[6] M. Cirstea and M. Levin, "Compensatory strategies for reaching in stroke.," Brain, vol. 123, pp. 940-953, 2000.

[7] M. F. Levin, "Interjoint coordination during pointing movements is disrupted in spastic hemiparesis.," Brain, vol. 119, pp. 281-293, 1996.

[8] T. Shaikh, V. Goussev, A. Feldman, and M. Levin, "Arm-trunk coordination for beyond-the-reach movements in adults with stroke.," Neural Repair, vol. 28, pp. 355-366, 2014.

[9] M. F. Levin, D. G. Liebermann, Y. Parmet, and S. Berman, "Compensatory versus noncompensatory shoulder movements used for reaching in stroke," Neurorehabilitation and neural repair, vol. 30, pp. 635-646, 2016.

[10] C. E. Lang, J. M. Wagner, A. J. Bastian, Q. Hu, D. F. Edwards, S. A. Sahrmann, et al., "Deficits in grasp versus reach during acute hemiparesis," Experimental Brain Research, vol. 166, pp. 126-1362005 ,.

[11] D. G. Liebermann, S. Berman, P. L. Weiss, and M. F. Levin, "Kinematics of reaching movements in a 2-D virtual environment in adults with and without stroke. ," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 20, pp. 778-787, 2012.

[12] R. Osu, K. Ota, T. Fujiwara, Y. Otaka, M. Kawato, and M. Liu, "Quantifying the quality of hand movement in stroke patients through three-dimensional curvature.," Journal of neuroengineering and rehabilitation, vol. 8, p. 62, 2011.

[13] J. Sanford, J. Moreland, L. R. Swanson, P. W. Stratford, and C. Gowland, "Reliability of the Fugl-Meyer assessment for testing motor performance in patients following stroke.," Physical therapy, vol. 73, pp. 447-454, 1993.

[14] G. Sprint, D. J. Cook, D. L. Weeks, and V. Borisov, "Predicting functional independence measure scores during rehabilitation with wearable inertial sensors," IEEE Access, vol. 3, pp. 1350-1366, 2015.

[15] D. J. Gladstone, C. J. Danells, and S. E. Black, "The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties. ," Neurorehabilitation and neural repair, vol. 16, pp. 232-240, 2002.

[16] S. H. Lee, M. Song, and J. Kim, "Towards clinically relevant automatic assessment of upper-limb motor function impairment," presented at the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), 2016.

[17] M. A. Villán-Villán, R. Pérez-Rodríguez, C. Gómez, E. Opisso, J. M. Tormos, J. Medina, et al., "Automated Fugl-Meyer assessment for ABI subjects in upper limb physical

" 2015.

[18] B. T. McMahon and L. R. Shaw, "advances in brain injury rehabilitation.," 1991.

[19] L. K. Kwah and R. D. Herbert, "Prediction of walking and arm recovery after stroke: a critical review.," Brain sciences, vol. 6, 2016.

[20] M. Mirbagheri and W. Z. Rymer, "Time-course of changes in arm impairment after stroke: variables predicting motor recovery over 12 months.," Archives of physical medicine and rehabilitation, vol. 89, pp. 1507-1513, 2008.

[21] P. Otten, J. Kim, and S. H. Son, "A framework to automate assessment of upper-limb motor function impairment: A feasibility study.," Sensors, vol. 15, pp. 20097-20114, 2015.

[22] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," ACM SigKDD Explorations Newsletter, vol. 12, pp. 74-82, 2011.

[23] P. Tormene, T. Giorgino, S. Quaglini, and M. Stefanelli, "Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation.," Artificial intelligence in medicine, vol. 45, pp. 11-34., 2009.

[24] R. Kizony, P. L. Weiss, O. Elion, S. Harel, I. Baum-Cohen, T. Krasovsky, et al., "Development and validation of tele-health system for stroke rehabilitation.," International Journal on Disability and Human Development, vol. 13, pp. 361-368, 2014.

[25] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview.," International Journal of Data Warehousing and Mining, pp. 1-13, 2007.

[26] L. Rokach, A. Schclar, and E. Itach, "Ensemble methods for multi-label classification," Expert Systems with Applications, vol. 41, pp. 7507-7523, 2014.

[27] M. Alaverdashvili and I. Whishaw, "A behavioral method for identifying recovery and compensation: Hand use in a preclinical stroke model using the single pellet reaching task. ," Neuroscience and Biobehavioral Reviews, vol. 37, pp. 950-967, 2013.

[28] P. W. McClure, L. A. Michener, B. J. Sennett, and A. R. Karduna, "Direct 3-dimensional measurement of

scapular kinematics during dynamic movements in vivo. ," Journal of shoulder and elbow surgery, vol. 10, pp. 269-277, 2001.

[29] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," arXiv preprint arXiv:1610.07717.2016 ,.

[30] A. J. Turton, P. Cunningham, E. Heron, F. van Wijck, C. Sackley, C. Rogers, et al., "Home-based reach-to-grasp training for people after stroke: study protocol for a feasibility randomized controlled trial.," Trials, vol. 14, p. 109, 2013.

[31] M. F. Levin, "Interjoint coordination during pointing movements is disrupted in spastic hemiparesis," Brain, vol. 119, pp. 281-293, 1996

[32] N. Hogan and D. Sternad, "Sensitivity of smoothness measures to movement duration, amplitude, and arrests," Journal of motor behavior, vol. 41, pp. 529-534, 2009.

[33] M. M.A., W. C., and S. K.S., "Kinematic Variables Quantifying Upper-Extremity Performance After Stroke During Reaching and Drinking From a Glass.," Neurorehabilitation and Neural Repair, vol. 25, pp. 71 –80, 2011.

[34] P. M. van Vliet and M. R. Sheridan, "Coordination between reaching and grasping in patients with hemiparesis and healthy subjects.," Archives of physical medicine and rehabilitation, vol. 88, pp. 1325-1331, 2007.

[35] M. C. Baniña, A. A. Mullick, B. J. McFadyen, and M. F. Levin, "Upper limb obstacle avoidance behavior in individuals with stroke.," Neurorehabilitation and neural repair, vol. 31, pp. 133-146, 2017.

[36] G. Doquire and M. Verleysen, "Feature selection for multi-label classification problems," presented at the In International work-conference on artificial neural networks, 2011.

[37] G. Holmes, A. Donkin, and I. H. Witten, "Weka: A machine learning workbench," 1994.

[38] G. Tsoumakas and I. Vlahavas, "Random k-labelsets: An ensemble method for multilabel classification.," presented at the In European conference on machine learning, 2007.

[39] G. C. Cawley and N. L. Talbot, "Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. ," Pattern Recognition, vol. 36, pp. 2585-2592, 2003.

[40] J. E. Staunton, D. K. Slonim, H. A. Coller, P. Tamayo, M. J. Angelo, J. Park, et al., "Chemosensitivity prediction by transcriptional profiling.," Proceedings of the National Academy of Sciences, vol. 98, pp. 10787-10792.2001 ,.

[41] M. L. Zhang and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern recognition, vol. 40, pp. 2038-2048, 2007.

[42] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization.," Machine learning, vol. 39, pp. 135-168, 2000.

[43] Y. Yang and X. Liu, "A re-examination of text categorization methods.," In Sigir, vol. 99, p. 99, 1999.

[44] W. W. Daniel, "Friedman two-way analysis of variance by ranks.," Applied nonparametric statistics, pp. 262-274, 1990.

[45] J. H. Morris, T. Oliver, T. Kroll, S. Joice, and B. Williams, "Physical activity participation in community dwelling stroke survivors: synergy and dissonance between motivation and capability. A qualitative study. ," Physiotherapy, vol. 103, pp. 311-321, 2017.

[46] C. M. Bishop, "Pattern recognition and machine learning. ," springer., 2006.

[47] J. Langan, K. DeLave, L. Phillips, P. Pangilinan, and S. H. Brown, "Home-based telerehabilitation shows improved upper limb function in adults with chronic stroke: a pilot study.," Journal of rehabilitation medicine., vol. 45, pp. 217-220, 2013.

[48] Eizicovits D., Edan Y., T. I., and L.-T. S., "Robotic gaming prototype for upper limb exercise: effects of age and embodiment on user preferences and movement. ," Restorative Neurology and Neuroscience, vol. 36, pp. 261–274, 2018.

[49] Feingold-Polak R., Elishay A., Shahar Y., Stein M., Edan Y., and L.-T. S, "Differences be-tween young and old users when interacting with a humanoid robot: a qualitative us-ability study.," Paladyn, Journal of Behavioral Robotics, vol. 9, pp. 183-192, 2018.

[50] Kashi S. and L.-T. S., "Smooth leader or sharp follower? Playing the mirror game with a robot.," Restorative Neurology and Neuroscience vol. 36, pp. 147-159 2018.

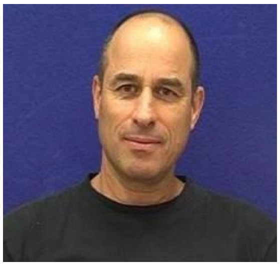
[51] A. Langer and S. Levy Tzedek, "Priming and Timing in Human-Robot Interactions," in Modelling human motion: from human perception to robot design, N. Noceti, A. Sciutti, and F. Rea, Eds., ed: Springer, 2020 (in press).

[52] Kellmeyer P., Müller O., F.-P. R., and L.-T. S., "Social Robots in Rehabilita-tion: A Question of Trust.," Science Robotics, vol. 3, p. 1587, 2018.

[53] A. Langer, Feingold-Polak R., Müller O., K. P., and L.-T. S., "Trust in Socially Assistive Robots: Considerations for use in Rehabilitation.," Neuroscience and Biobehavioral Reviews, vol. 104, pp. 231-239, 2019.

Shir Kashi is a master's degree student in Software and Information Systems Engineering at Ben-Gurion University of the Negev (BGU). She completed her bachelor’s degree in neuroscience and computer science in 2017 at BGU. She won the best late-breaking report award at the 2017 Conference on Human-Robot Interaction (HRI) in Vienna, Austria for her work with a social robot for rehabilitation.

Ronit Feingold-Polak receveid her Bachelor of Physical Therapy degree 2002, and her MScPT degree in 2011. She has worked as a Physical Therapist for the last 17 years in several rehabilitation units, including the Sheba Medical Center in Israel. She is specialized in treating acquired brain injury rehabilitation. Currently pursuing her PhD in Physical Therapy at the Ben-Gurion University of the Negev. Her research focuses on long-term interaction of post-stroke survivors with socially assistive robots. She is the co-author on three journal papers, has presented her work at several international conferences, and is a student member of the ACM IEEE.



Boaz Lerner received a B.A. degree in Physics and Mathematics from the Hebrew University, Israel, in 1982 and a Ph.D. degree in Computer Engineering from Ben-Gurion University, Israel, in 1996. He was a researcher at the Neural Computing Research Group at Aston University, Birmingham, UK and the Computer Laboratory of Cambridge University, Cambridge, UK, and now is an Associate Professor in the Department of Industrial Engineering and Management at Ben-Gurion University, Israel, where he established in 2007 and has since headed the Machine Learning and Data Mining Lab. His current interests include machine learning and data mining approaches to data analysis and their application to real-world problems. Lerner has supervised nearly 50 graduate students and has published around 100 papers in journals and conference proceedings.

Lior Rokach is a data scientist and a Professor of Software and Information Systems Engineering (SISE) at Ben-Gurion University of the Negev (BGU). He is also the current chair of the department. His research interests lie in the design and analysis of Machine Learning and Data Mining algorithms and their applications in Recommender Systems, Cyber Security and Medical Informatics. Prof. Rokach has established the machine learning lab at BGU which promotes innovative adaptations of machine learning and data science methods to create the next generation of intelligent systems. Prof. Rokach is the author of over 300 peer-reviewed papers in leading journals and conference proceedings, patents, and book chapters.

Shelly Levy-Tzedek obtained her B.Sc. degree summa cum laude in Bioengineering from the University of California, Berkeley in 2002, and her M.Sc and Ph.D. degrees in Biological Engineering from the Massachusetts Institute of Technology (MIT) in 2004 and 2008, respectively. She heads the Cognition, Aging and Rehabilitation laboratory at the Department of Physical Therapy at the Ben-Gurion University of the Negev, where she is also a member of the Zlotowski Center for Neuroscience and the ABC robotics initiative. In 2018-2019 she was awarded a Horizon-2020 Marie Skłodowska Curie Visiting Professorship at the FRIAS Institute of Advanced Studies at the University of Freiburg in Germany. She has been awarded the Pedagogica award for outstanding new researchers in 2016 and the Toronto Prize for excellence in research in 2018. Her work is funded by national and international foundations - both public and private.