Monitoring Arm Movements Post-Stroke for Applications in Rehabilitation and Home Settings

Juan Pablo Gomez-Arrunategui, Janice J. Eng, and Antony J. Hodgson

***Abstract*—Optimal recovery of arm function following stroke requires patients to perform a large number of functional arm movements in clinical therapy sessions, as well as at home. Technology to monitor adherence to this activity would be helpful to patients and clinicians. Current approaches to monitoring arm movements are limited because of challenges in distinguishing between functional and non-functional movements. Here, we present an Arm Rehabilitation Monitor (ARM), a device intended to make such measurements in an unobtrusive manner. The ARM device is based on a single Inertial Measurement Unit (IMU) worn on the wrist and uses machine learning techniques to interpret the resulting signals. We characterized the ability of the ARM to detect reaching actions in a functional assessment dataset (functional assessment tasks) and an Activities-of-Daily-Living (ADL) dataset (pizza-making and walking task) from 12 participants with stroke. The Convolutional Neural Network (CNN) and Random Forests (RF) classifiers had a Matthews Correlation Coefficient score of 0.59 and 0.58 when trained and tested on the functional dataset, 0.50 and 0.49 when trained and tested on the ADL dataset, and 0.37 and 0.36 when trained on the functional dataset and tested on the ADL dataset, respectively. The latter is the most relevant scenario for the intended application of training during a clinical visit for monitoring movements in the in-home setting. The classifiers showed good performance in estimating the time spent reaching and number of reaching gestures and showed low sensitivity to irrelevant arm movements produced during walking. We conclude that the ARM has sufficient accuracy and robustness to merit being used in preliminary studies to monitor arm activity in rehabilitation or home applications.**

***Index Terms*—Accelerometer, home monitoring, arm rehabilitation, reach detection, machine learning.**

I. Introduction

Research into recovery of arm function following stroke has established that repeated voluntary movements lead to improvements in arm function [1] by inducing neural growth and brain reorganization [2]. However, performing rehabilitation exercises in the clinic alone cannot achieve sufficient repetitions for optimal recovery given that the typical hourly clinical session averages only 32 arm and hand repetitions [3].

Therapists have therefore sought to encourage in-home rehabilitation exercise programs to supplement functional recovery but find it challenging to monitor patients' adherence to prescribed therapeutic protocols and to assess the quality of exercises when not supervised by a therapist. Research studies have used commercial wrist-mounted accelerometers to quantify arm movement in home settings, primarily in the form of activity counts [4]. Activity counts record the occurrence of movement, but do not otherwise assess the context, purpose or quality of the movement [5]. Therapists are most concerned with knowing how often patients perform functionally meaningful arm movements, rather than incidental movements such as the arm swing that occurs with walking; such incidental movements should ideally be excluded from automatically documented movement histories. Researchers have tried to overcome these limitations by asking participants to wear additional sensors to detect walking, and then removing these time periods [6]. However, this approach burdens participants with additional data logging expectations and burdens researchers with post-processing of the activity counts.

A more recent implementation of activity counts proposed by Bailey *et al.* [7] compared the relative activity counts of accelerometers on both the paretic and non-paretic limb to provide information on intensity of bilateral activity and the contribution of each limb to the activity. Metrics obtained from this approach can distinguish the intensity of tasks and whether tasks were completed with both arms but are still unable to provide information on functional arm use.

In recent years, classifiers have been used to develop gesture recognition systems that can be used for monitoring and classifying movements. Such developments have mostly occurred in the lab setting, where there are many measurement resources and greater experimental control over participant gestures. Activity recognition in the lab can use multiple sensors attached to the arms, legs and torso to detect limb orientations and movements, which can substantially increase the detection accuracy relative to using single sensors, though the increased complexity of multiple-sensor setups can compromise user comfort and present barriers to use in the home setting [8], [9].

Manuscript received 20 January 2022; revised 7 June 2022 and 3 August 2022; accepted 7 August 2022. Date of publication 10 August 2022; date of current version 22 August 2022. This work was supported in part by the Canadian Institute of Health Research under Grant MOP-137053, in part by the Natural Sciences and Engineering Research Council under Grant CHRP 462256-2014, and in part by the Canada Research Chair Programs. *(Corresponding author: Juan Pablo Gomez-Arrunategui.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the UBC Behavioural Research Ethics Board under Application No. H15-02613 and Operational Approval was granted by the Vancouver Coastal Health Authority under Application No. V15-02613.

Juan Pablo Gomez-Arrunategui is with the Department of Mechanical Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: jpgarrunategui@gmail.com).

Janice J. Eng is with the Department of Physical Therapy, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

Antony J. Hodgson is with the Department of Mechanical Engineering and the School of Biomedical Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

Digital Object Identifier 10.1109/TNSRE.2022.3197993

In addition to the variety of hardware configurations that have been used for this purpose, there have also been many machine learning approaches applied to gesture recognition problems, ranging from k-nearest neighbor with similarity measures [10], [11], to Dynamic Bayesian Models [10], Hidden Markov Models [12], [13], Support Vector Machines [8], [14], and ensemble methods such as Joint Boosting [8], [15] or Random Forests [16], [17]. More recent literature has explored convolutional neural networks, which have been increasingly successful in image recognition tasks, and have the added advantage of automating feature learning from raw data inputs, eliminating the need for hand-tailoring features for task recognition [18], [19], [20], [21], [22].

Regardless of the preferred classifier, most machine learning approaches train and test their classifiers on a controlled set of activities. This constrains the classification to a reduced number of standard tasks and typically ignores the existence of an undefined set of activities that the classifier has not been trained to detect. This is particularly relevant to a device intended for home monitoring purposes, as the system needs to be able to recognize the wide range of arm movements that are performed in daily life that are nonetheless not meaningful from the perspective of promoting functional recovery.

Lum *et al.* [23] addressed this issue by exploring a gesture classification system that detects functional activity in stroke participants outside of the lab setting. Their approach, consisting of a single accelerometer on the paretic limb, showed that machine learning algorithms can provide more accurate information of functional activity than traditional methods that use accelerometer counts. They specifically showed improved detection of functional activity when compared against counts ratio (a method to calculate duration of activity in the paretic limb normalized by the less-affected limb), which did not significantly correlate with functional movement.

In this paper, we propose and evaluate a method intended for use in recognizing functional “reaching” actions for home monitoring applications that must account for the existence of a large undefined set of gestures. By focusing initially on reaching actions – i.e., a certain class of fundamental movements that form the base for more complex gestures and actions such as grasping and object manipulation – we hope to demonstrate the ability to reliably detect a significant subset of functional gestures performed in unconstrained (home-like) settings. We call our proposed system the **Arm Rehabilitation Monitor (ARM)**. With ARM, we acquire arm movement information using a single wrist-mounted inertial measurement unit (IMU) sensor (in order to minimize obtrusiveness to the participant and thereby increase adherence to using the monitor) and process the data with machine learning classifiers to discriminate between reach and non-reach actions, where reach is defined as a movement of the shoulder and elbow away from the body, and non-reach actions encompass any possible gestures other than reach.

II. Methods

We developed an apparatus and method for acquiring and processing IMU data and conducted a user study involving participants living with a stroke who were asked to perform both standard reach assessment tasks in the clinic and tasks simulating a complex Activity of Daily Living (ADL), similar to what they might perform in a home setting. We report the discriminative accuracy of the developed method on these tasks.

A. Participants

For the user study, we recruited 12 participants. The mean age of the participants (5 women and 7 men) was 65.4 ± 13.0 and the time since stroke was 10.0 ± 7.2 years. The mean Fugl Meyer score was 57.3 ± 10.0, and the mean ARAT score was 49 ± 11.7, which indicates mild to moderate hemiparesis in the upper extremities. The ARAT score was later used to determine whether our classification algorithms’ accuracy differs between participants with varying levels of impairment. This study received institutional review board approval from the UBC Behavioural Research Ethics Board (H15-02613) and operational approval from the Vancouver Coastal Health Authority (V15-02613).

B. User Study

We asked the participants to perform two assessments in the laboratory while wearing the ARM sensor (described below) on the wrist of the most-affected limb to generate two datasets: (1) a functional assessment dataset that consists of a series of standard functional assessment movements and (2) an ADL dataset that consists of a pizza-making task, the latter supplemented by a walking task aimed at collecting arm movement data during a confounding task (i.e., not involving a purposeful reach gesture). All tasks were video recorded.

We used arm movement data obtained by the ARM sensor to train a gesture discriminator (classifiers described below) and subsequently tested the discriminator’s accuracy in identifying reach gestures. We characterized the ability of the gesture discriminator to learn on the two different datasets as well as its ability to transfer learning from one dataset to another. The walking task was used to estimate the ability of the discriminator to distinguish between intentional reaching movements and incidental limb movements due to walking.

Additionally, during the ADL task, we fitted all the participants with a commercially available activity count monitor (Actical™, MM; Mini-Mitter Co) on the same wrist as the ARM to compare the ARM’s ability to detect reach gestures with the Actical’s activity counts. Once the study had started, we realized it would be beneficial to also record Actical data during the functional assessment task; we therefore also present Actical data for the functional assessment task for the final 6 participants.

C. Data Acquisition

The ARM monitor contains a TDK InvenSense® 9-axis IMU (MPU9150) attached to the wrist by a watch strap (see Fig. 1). The ARM is placed on the wrist of the most-affected arm in the same position as a watch. The

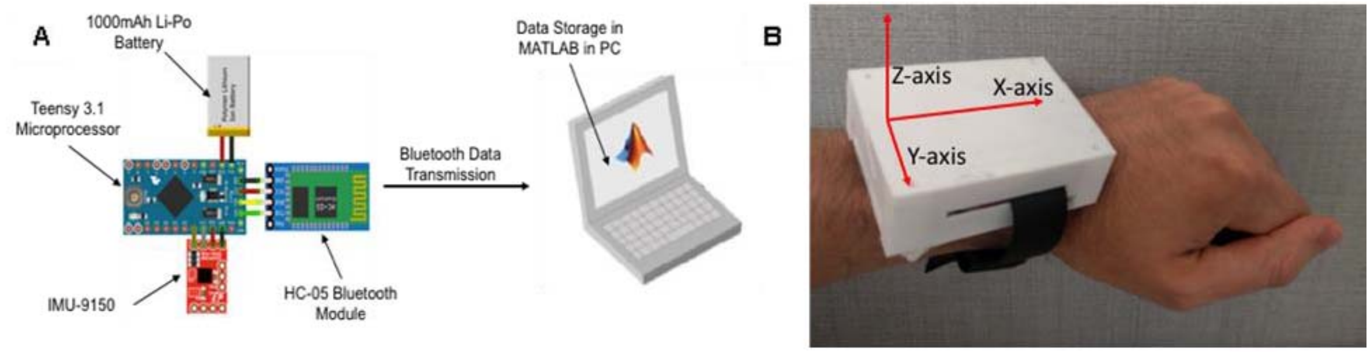

A
1000mAh Li-Po Battery
Teensy 3.1 Microprocessor
IMU-9150
HC-05 Bluetooth Module
Bluetooth Data Transmission
Data Storage in MATLAB in PC
B
Z-axis
X-axis
Y-axis

Fig. 1. A. System diagram depicting the IMU for data acquisition, the Teensy 3.1 microprocessor, the HC-05 Bluetooth module for data transmission, the battery that powers the wrist-worn device and the PC that stores the data for offline analysis. B. Picture of the ARM with the IMU coordinate system.

IMU produces a data stream consisting of three orthogonal acceleration channels, three rotational rate channels, and a three-axis digital compass. Only the orthogonal acceleration channels and rotational rate channels were used for gesture discrimination because the digital compass channels were found in pilot studies to provide less relevant information for reach gesture recognition. The IMU acquired movement data at a sampling rate of 20 Hz, which is sufficient for arm monitoring [24]. The data were transmitted over Bluetooth to a computer, where it was stored for labelling and offline gesture recognition.

The tasks were also video recorded with a video camera (30 frames per second). The video camera provided a visual record of the hand activities and the video footage was later used to manually annotate the data.

The Actical sensor mentioned above was also attached on the wrist of the most affected arm and acquired data at a sampling rate of 32 Hz. The Actical data was integrated every 15 seconds to generate a metric of activity counts. The activity counts were stored in the Actical monitor during the tests. The data were downloaded to a computer and stored when the user study was complete.

D. Datasets

Data was collected during the functional assessment and during an ADL task. The functional assessment dataset was composed of measurements made during functional assessment tasks aimed at determining the functional level in the most-affected arm. We opted to use the motor function score for the upper extremities of the Fugl-Meyer Assessment, which has 22 upper extremity movements, and the "light touch" section of the sensory function assessment, which has two tasks [25]. We also used the Action Research Arm Test (ARAT), which has 19 upper extremity movements [25], to evaluate hand sensorimotor function. The Fugl-Meyer and ARAT are two of the recommended key measures of the upper extremity as determined by an international consensus panel on stroke recovery [26]. Participants were asked to complete each of the tasks in the assessments once. Reach gestures made while performing these two functional assessments were used to train and evaluate the reach discrimination models.

TABLE I
Activities and Corresponding Action Label

Activity	Action Label
Reach for roller	Reach
Roll dough	Reach
Reach out for pizza sauce	Reach
Open can	Grab/rotate (non-reach)
Reach for spoon	Reach
Pour sauce on pizza	Wrist rotation (non-reach)
Reach for ingredient	Reach
Place ingredient on pizza	Reach

On average, participants performed 65 reach gestures in the functional assessment dataset and spent 20 minutes completing the assessments.

The ADL dataset was composed of tasks in which the participant performs a complex activity of daily living. We chose to have participants make pizzas because performing this task requires a wide range of natural (non-stereotyped and non-repetitive) reaching movements and because participants require little explicit instruction. This leads to participants performing a more natural set of movements that better mimic their home environment. We anticipated significant variability across participants in the number and order of reaching actions.

The pizza-making task had three primary steps:

1. Roll the dough
2. Add pizza sauce to the dough
3. Populate the pizza with any combination of the ingredients provided

There were no instructions regarding the participant's body position when making the pizza. Nine participants decided to remain seated, while the other three preferred to stand up. The activity took approximately fifteen minutes and participants spent an average of 28.1% of their time performing functional arm movements (reaching actions). The range of participant activities observed is summarized in Table I above.

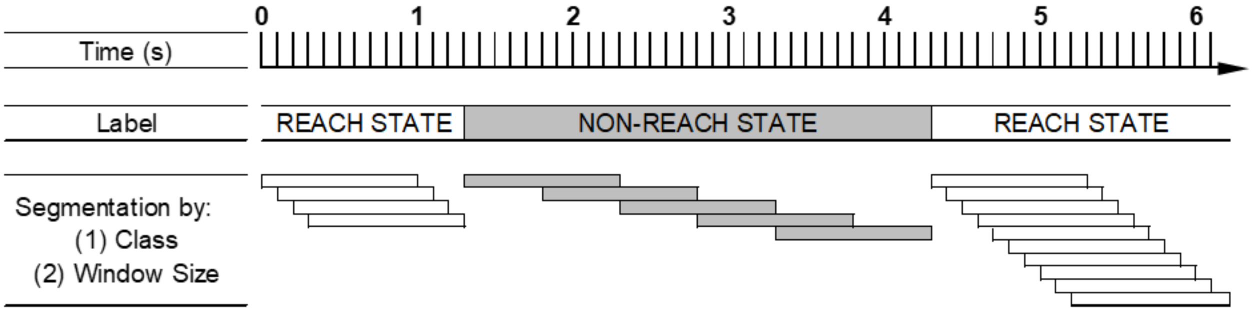


Fig. 2. Segmentation of training and validation data by: (1) Class, and (2) window size. To ensure that all windowed samples have unambiguous labels, all windowed samples are fully contained within one interval with the same ground-truth label (i.e., no samples cross a label boundary).

After completing the pizza task, eight participants with the ability to walk without external assistance (such as the use of a cane) were asked to walk around the room. These participants walked continuously for approximately 3 minutes while being recorded with both the Actical accelerometer and the ARM.

E. Data Labelling

To label the data, we classified all movements as either arm reaches or non-reaches. We defined an arm reach as any goal-directed movement of the arm moving away from the body. To ensure consistency, we required that the reach must involve “visible” movement at the elbow and shoulder.

This definition of reach was used to label the datasets through video recordings. One individual (JPG) manually labeled all reaches from the datasets using a video labeling tool that enabled the recording to be divided into episodes of reach and non-reach so that the durations of each reach and non-reach could also be identified. As the analyst gained experience in labeling the movements, we realized that arm reach cycles are complex movements that can typically be decomposed into three primary sub-events:

1. **Forward movement of arm:** Arm moves away from body; typically evidenced by movements at the elbow and shoulder
2. **Stationary (optional):** Pause to collect object. This phase may vary in length or may even not be detectable (e.g., in situations in which this event is too short to be recognized, such as when there is no object to be collected).
3. **Backward movement of the arm:** Arm returns to body; again, typically evidenced by simultaneous movements of the elbow and shoulder.

An arm reach cycle is therefore normally composed of two dynamic events (forward and backward movements), often with a static event (pause) in between. The functional nature of the arm reach is related to the dynamic events, which require movement of the elbows and shoulders. The pause in the middle of the reach may also be functional if it involves grasping an object. However, our current study does not include grasp detection, so, for our purposes, the pause corresponds to an event that we do not wish to detect, and we therefore labelled it a non-reach.

F. Data Processing

Before being used by the gesture discriminator (either for training or discrimination), the data channels were first preprocessed to attenuate high frequency noise. We applied an acausal low pass 3rd order Butterworth filter at 6 Hz. A similar filter was used by Biswas *et al.* [14] to filter high frequency noise artifacts and was found to work empirically.

The training data were originally segmented by class, per the ground-truth labels assigned through video-labelling. This class-segmented data contains examples of reach and non-reach movements of varying lengths, per the amount of time spent by the user in each of these states. The data were then segmented with 1.0s sliding windows to generate uniform length data fragments for use by the gesture discriminator (see Fig. 2). 1.0s sliding windows were preferred over longer window sizes (1.5s and 2.0s were also explored) because they proved in pilot work to be better able to distinguish individual reach instances while also being long enough to capture meaningful characteristics of reach movements. For training and validation, the overlap between consecutive windows was varied depending on the label assigned to the window. If the window was labelled as a non-reach, there was 50% overlap between consecutive windows. If the window was labelled as reach, there was 90% overlap between consecutive windows. Oversampling the reach class was done to reduce the class imbalance in the available datasets. The time spent in non-reach states in the functional assessment dataset was approximately ten times longer than the time spent in a reach state, whereas the non-reach time in the ADL dataset was about four times longer than the reach time.

The experimenter used a video annotation tool to generate the ‘ground truth’ reference for the testing data. The testing data was then segmented into window sizes that could be input into the classifier. The label of the testing data was unknown to the classifier. As such, the testing data was not segmented by class and individual data windows could contain labels for multiple classes. If a segmented window contained two different labels (reach and non-reach), the window was assigned the label that was present more often. In cases where the segmented window has an equal amount of reach and non-reach labels, the window was given the label of the previous window. The overlap between consecutive windows was set at 90% (see Fig. 3).

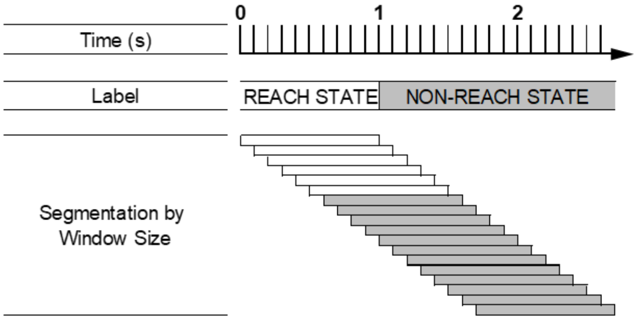


Fig. 3. Segmentation of testing data exclusively by window size without reference to underlying ‘ground truth’ labels.

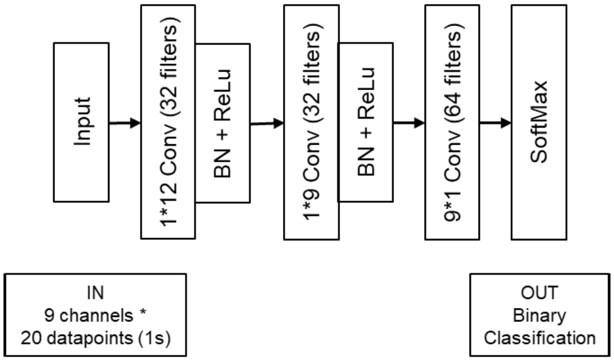


Fig. 4. CNN Architecture 1.0s window-segmented data.

The training data were normalized to zero mean and unit standard deviation. The validation and testing data were then normalized with the same parameters as the training data. The effect of standardizing the data was two-fold:

a. It equalized the range of the data in the different channels, which allowed the training classifiers to assign equal importance to the multiple data streams [27]
b. It adjusted the baseline for each user and/or each activity performed by that user.

After classification by the gesture discriminator, we assigned a label to each sampling point by treating the label associated with each window segment overlapping that sampling point as a ‘vote’ for that label and then taking the label with the highest number of votes (majority voting).

G. Classifiers for Gesture Recognition

We chose to test two classifiers for gesture recognition: (1) Random Forests (RF) and (2) Convolutional Neural Networks (CNNs). The RF classifier is an ensemble classifier that classifies data based on hand-tailored features. This classifier employs a collection of decision trees to classify data. Each decision tree is trained on a different sample of the data to gain unique insights into the data structure. Individual decision trees are likely to overfit the presented data, so the RF averages out the output from each tree to improve the classification accuracy. CNNs were chosen because they can be applied without predefining or hand-engineering the features used; in effect, they learn to recognize relevant features automatically [19]. Each convolutional layer in the CNN performs a non-linear data transformation that allows it to learn complex features to classify gestures from a multi-variate time-series data stream.

1) Random Forests: Random Forests are robust and relatively easy to tune when compared to other classification classifiers that employ hand-picked features such as Naïve-Bayes and Support Vector Machines [28]. They are an ensemble-type classifier (classifiers that average many individual results) that have previously been shown to produce good results in human activity recognition tasks [29].

We selected seven signal-based features (described below) for each of nine data channels (the gyroscope data in X, Y, and Z, accelerometer data in X, Y, Z, and the roll, pitch, yaw calculated by the IMU from the gyroscope and accelerometer data), and six correlation-based features (described below) that, based on initial pilot testing, showed good discrimination potential for reach detection. The signal-based features were extracted from each channel in the window-segmented data. The correlation-based features measured changes across multiple channels. The utility of pairwise correlations as features to detect activities in multiple dimensions has been previously demonstrated [15].

The 7 signal-based features we used were: (1) the mean of the signal, (2) the variance of the signal, (3) the root mean square (RMS) of the signal (the RMS is square root of the signal power), (4) the minimum value in the signal, (5) the maximum value in the signal, (6) the skewness of the signal, which is a measure of symmetry in the data (the formula for skewness is $\frac{\sum_{i=1}^{n}(X_i-\mu)^3/N}{\sigma^3}$) and (7) the kurtosis of the signal (a measure of the distribution of the data). A high kurtosis means that the data has many outliers relative to the normal distribution. The formula for kurtosis is $\frac{\sum_{i=1}^{n}(X_i-\mu)^4/N}{\sigma^4}$. The signal-based features were calculated for the 9 different channels described above, so the total number of signal-based features is therefore 63 (9 channels × 7 features per channel). We computed the following six Pearson correlations: 3 acceleration correlations (X vs Y, X vs Z and Y vs Z) and 3 gyroscope correlations (X vs Y, X vs Z and Y vs Z).

We set the number of features to select at random for each decision split to be the square root of the number of features. The leaves of the trees were not pruned so that the decision trees generated a greater number of decisions to use in learning the data. Based on a preliminary pilot study, we selected 50 as the number of trees.

2) CNN: We opted to design a CNN similar to Wang’s [21] FCN model, (see Fig. 4) with three blocks in the entire architecture. The input to the CNN is a 1 second window (20 datapoints) of the 9 channels of data (gyroscope data in X, Y, and Z, accelerometer data in X, Y, Z, and the roll, pitch, yaw calculated by the IMU from the gyroscope and accelerometer data) generated by the ARM monitor. The output of the CNN is a binary classifier that assigns a label (reach or non-reach) to each windowed segment of data.

The basic block is a convolutional layer followed by a batch normalization (BN) layer and a ReLU activation layer. The convolution operation of the first two blocks is done across the temporal space, while the third convolution operation is done across the channels. The first convolution operation is done by a 1-D kernel of size 1(channels)*12(datapoints) and a filter size of 32. The second convolution operation is done by a 1-D kernel of size 1(channels)*9(datapoints) and a filter size of 32. The third convolution operation is done by a 1-D kernel of size 9(channels)*1(datapoint) and a filter size of 64. The first two convolution sizes were large to capture more temporal information. The number of filters was reduced with respect to Wang's [21] model to decrease the number of hyperparameters to be learned given the small size of the dataset, which consists of an average of 2125 training instances per participant. Batch normalization is applied to speed up the convergence and improve generalization [21].

The CNN uses a learning rate of 0.001, a batch size of 32, trains over 16 epochs, and uses categorical cross entropy as the loss function.

A validation dataset was generated by splitting 20% of a continuous section of the training dataset from each participant at a randomized location. The randomized split of the training dataset was done to ensure different sections of the dataset were captured in the validation dataset, such that it would provide adequate information of a classifier's ability to generalize to the test dataset. The classifier hyperparameters were tuned with the results from the validation dataset.

H. Evaluation of Classifier Performance on Datasets

The functional assessment dataset and the ADL dataset represent, respectively, data acquired from more and less constrained functional tasks. The functional assessment dataset represents data that is more readily obtainable in the clinical setting, as it is comprised of testing protocols that are commonly administered there. In contrast, the ADL dataset is a proxy for the less constrained tasks normally performed in the home setting, which is our desired site of application. We designed three experiments to evaluate three potential future applications of the classifier.

1) Clinic-Clinic (Train on Functional Assessment Dataset, Test on Functional Assessment Dataset): This experiment measures how well the classifiers are able to learn to recognize gestures acquired in the course of a standardized assessment in the clinical environment, where it is easier to collect data from participants, and generalize to a new participant, also assessed in the clinic. The classifiers are trained on functional assessment data from all users but one in our study and are then tested on the functional assessment data from the user who was left out from the training phase. This is repeated for all the participants in the dataset. This would be of modest clinical utility since the primary purpose is to assess patient movements in the home environment, but because both the training and tests assessments use standardized protocols, the classifiers' performance is likely to be best in this situation.

2) Home-Home (Train on ADL Dataset, Test on ADL Dataset): This experiment measures how well the classifiers are able to learn from training on one group of participants in the home environment and generalize to a new participant in their home environment. For this application, the ADL dataset is used as the proxy for the home setting. This ADL dataset captures the tradeoff between data richness and the capacity to discriminate reach gestures in more unconstrained environments. It is more difficult and time consuming to label the ADL dataset than the functional assessment dataset because arm movements are not as stereotyped as they are in the assessment process. The classifiers are trained on the ADL dataset from all participants but one and are then tested on the ADL dataset from the participant the was left out from the training phase. This is repeated for all the participants in the dataset. This type of classifier would potentially be of high clinical value as it would enable use in the home environment, but it would be difficult to train such a classifier on movements acquired in patients' homes due to the difficulty of obtaining reference data in this setting. Nonetheless, this experiment will enable us to estimate how well such a classifier could potentially perform.

3) Clinic-Home (Train on Functional Assessment, Test on ADL): This experiment assesses how well the classifiers are able to learn in the more practically realistic setting in which training data is obtained on more constrained tasks in the functional assessment and subsequently applied to less constrained tasks in the home setting. Here, the classifiers are trained to detect reach gestures on the data from the functional assessment dataset and are subsequently tested on the ADL dataset. This experiment measures the performance we might expect from training in the clinic (the functional assessment dataset) using data from all participants (including the target subject, since they would have a clinical evaluation before being sent home for monitoring) and applying them to the home setting (modelled by the ADL dataset for the target). This configuration most closely resembles the target application of a home monitoring system based on training data acquired in clinical evaluations.

I. Evaluation Metrics

The accuracy, precision, recall, and Matthew's correlation coefficient (MCC) were used to evaluate classifier performance. The accuracy is used since it is a recognizable metric that can be used to compare our results to other studies. The other metrics are used because they give a better indication of performance in imbalanced datasets, like the ones used in this study. Precision is the share of predicted reach instances that are correct. Recall is the share of actual reach instances that are predicted correctly. The MCC will generate a high score if the binary predictor is able to correctly predict both the majority of positive data instances (reaches) and the majority of negative data instances (non-reaches). The MCC will lie in the interval [−1, +1], with the extreme values of −1 and +1 reached in cases of perfect misclassification and perfect classification, respectively, while MCC=0 is the expected value for a random classifier.

Precision, recall, and MCC are calculated as follows:

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where, TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

We also generated two clinically relevant metrics of classifier performance on the experiments by estimating: **(1) Reach time**, and **(2) Reach counts**. The **reach time** is the sum of time durations for all sampling instances which are classified as reaches. The **reach count** is the sum of the number of instances where there are consecutive sampling instances labelled as reaches with non-reaches on either side. Discrete reach counts are relatively easy to understand, provide the participant with quantitative data on arm activity, and are superior to the activity counts provided by most existing commercial activity trackers, which do not have a meaningful measurement unit. These two metrics reflect related, yet distinct, considerations. For example, if a reach occurred that lasted 1.0 s as estimated from the video record, and our device properly identified that a reach occurred, but estimated its duration as being only 0.75 s, then the reach time accuracy would be 75%, but the reach count accuracy would be 100%.

III. Results

Table II. summarizes the performance of the classifiers on the three experiments. The CNN and RF classifiers show similar performance, with no statistically significant differences on the accuracy or MCC scores between the two classifier types on the different experiments when evaluated with a paired t-test. Both classifiers showed the best accuracy and MCC scores in the clinic-clinic configuration, followed by the home-home configuration, and finally the clinic-home configuration.

A. Reach Time Prediction

In order to evaluate the ability of the classifiers to estimate the time participants spent performing reaching actions, we plotted the cumulative reach time predicted by each classifier against the cumulative reach time, as labelled in the video recordings, at each instance in the dataset (see Fig. 5. A below). The plot line therefore starts at time=0, when the cumulative predicted reach time is 0 s and the cumulative labelled reach time is also 0 s. The endpoint of the plot line shows the cumulative predicted reach time against the labelled reach time at time=end (over the entire dataset). If the resulting plot line rises at 45 degrees, this means that the classifier perfectly predicts the duration and time at which a reaching action occurs. Therefore, a plot line with an angle above 45 degrees is over-predicting the amount of time spent reaching at that moment, while any line with an angle under 45 degrees is under-predicting the amount of time spent reaching at that moment. The results of Fig. 5. A do not show a strong dependency of reach time prediction errors on the degree of impairment of the participant for either classifier.

Fig 5. B compares the % time spent performing reaching actions in the test dataset (as measured by video) against the classifier predictions. An ANOVA, followed by t-tests with a Bonferroni correction reported a significant difference in the reach time % between the video labelling and the CNN classifier for the Clinic-Home experiment (P < .005).

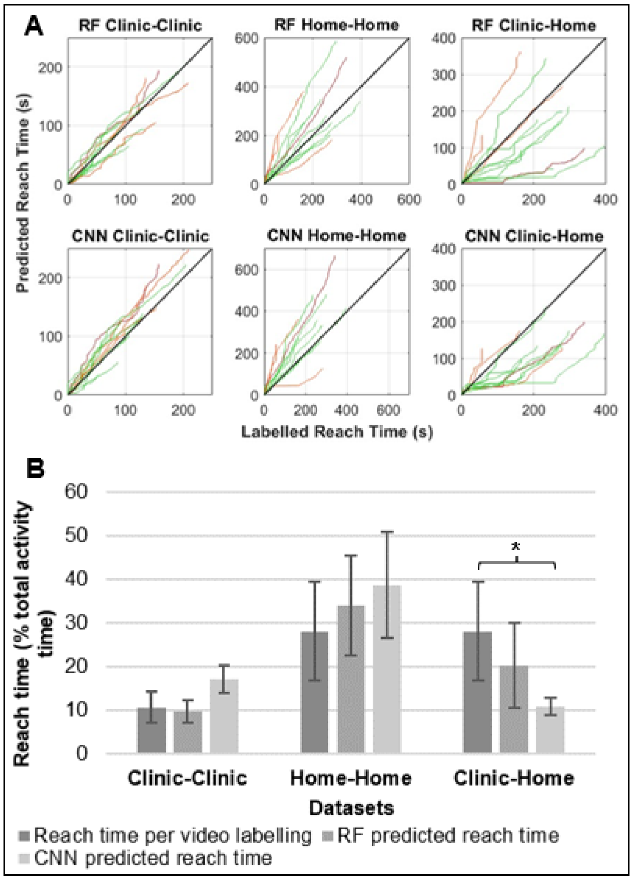


Fig. 5. A. Cumulative predicted reach time (s) vs cumulative labelled reach time (s) for all test configurations and both classifier types. Each participant is plotted as a different line, and the degree of impairment for the most affected arm is colour-coded (green: score >= 50 on ARAT scale, orange: 30 >= ARAT < 50, and red: ARAT < 30). B. Reach time (per video labelling) compared against the predicted reach time by RF and CNN classifiers. *Significant difference, p < .05.

B. Reach Count Prediction

Similarly, we evaluated the ability of the classifiers to estimate reach counts. Table III shows the precision and recall of the classifiers in estimating the number of reaching counts Note that, as described in the evaluation metrics section, a successfully predicted reach count occurs whenever there is overlap between the classifier prediction and the video-label, regardless of whether the start point and end points of the reaching action match. As such, classifiers are more likely to correctly predict reach counts than reach time, which leads to higher precision and recall scores in Table III than Table II.

Fig. 6 below compares the classifiers' cumulative predictions of reach counts to the number of reach counts identified on the video recording at each point in time. This plot shows the same general trends as for the reach times: generally tighter correspondence in slope in the Clinic-Clinic mode and more varied in the Clinic-Home mode, few differences between the classifier types, and no obvious relationships between degree of impairment and slope.

Fig. 7 illustrates the cumulative Actical activity counts against the labelled reach counts for the functional assessment and ADL datasets. Once again, we plotted the activity counts

TABLE II
CLASSIFIER RESULTS

Classifier	Evaluation Metric	Experiments		
		Clinic-Clinic	Home-Home	Clinic-Home
Random Forests	Accuracy (%)	92.4 ± 1.8	77.1 ± 5.7	74.8 ± 9.5
	Precision (%)	64.8 ± 12.2	59.0 ± 17.4	58.8 ± 17.2
	Recall (%)	60.3 ± 11.1	72.8 ± 13.4	46.9 ± 24.8
	MCC (-1, 1)	0.58 ± 0.09	0.49 ± 0.05	0.36 ± 0.17
CNN	Accuracy (%)	92.2 ± 1.7	76.6 ± 5.0	76.5 ± 8.9
	Precision (%)	61.9 ± 10.6	56.4 ± 16.6	62.9 ± 14.0
	Recall (%)	66.2 ± 13.3	79.3 ± 16.2	43.0 ± 15.2
	MCC (-1, 1)	0.59 ± 0.09	0.50 ± 0.07	0.37 ± 0.13

TABLE III
CLASSIFIER RESULTS FOR REACH COUNT PREDICTIONS

Classifier	Evaluation Metric	Experiments		
		Clinic-Clinic	Home-Home	Clinic-Home
Random Forests	Precision (%)	66.5 ± 15.8	69.4 ± 17.0	67.9 ± 12.2
	Recall (%)	73.6 ± 23.9	93.3 ± 6.2	84.7 ± 10.5
CNN	Precision (%)	69.0 ± 13.3	63.4 ± 17.0	64.6 ± 10.7
	Recall (%)	74.2 ± 14.1	94.2 ± 8.2	88.7 ± 15.2

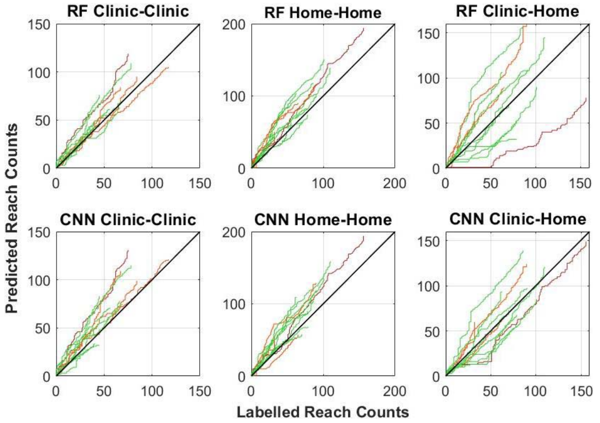


Fig. 6. Progression of predicted reach counts vs labelled reach counts for all test configurations.

at every point in time against the labelled reach counts at that moment. Note that because the Actical device measures an abstract activity metric called a ‘count’ that does not correspond to a discrete action such as a reach, we cannot directly compare the ARM reach count to the Actical activity count. Nonetheless, the lines in the functional dataset appear to have a similar relative range of slopes as we found with our classifiers, whereas the lines in the ADL dataset appear to have a larger range of slopes than our classifiers.

Participants that included walking activities in their ADL task showed a sudden increase in activity counts when walking was taking place. In fact, while walking tasks accounted for only 6.9% of the total time in the ADL dataset, on average

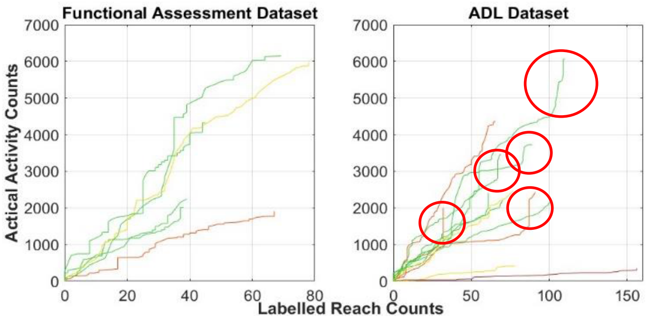


Fig. 7. Cumulative Actical activity counts vs video-labelled reach counts for the functional assessment dataset (left) and ADL dataset (right). The red circles on the right plot show the increase in Actical activity counts during walking when there are no functionally meaningful reaches occurring.

15.6% of the counts measured by the Actical happened during this time. This suggests that the Actical sensor confounds arm movements arising from walking with more specific reaching actions. In contrast, the ARM classifiers do not show this rapid increase in reach counts during walking activities. While there are incorrect reach predictions during walking, on average 5.3% of all reaches predicted by the classifiers took place during walking activities. This is roughly three times less than the proportion for the Actical system.

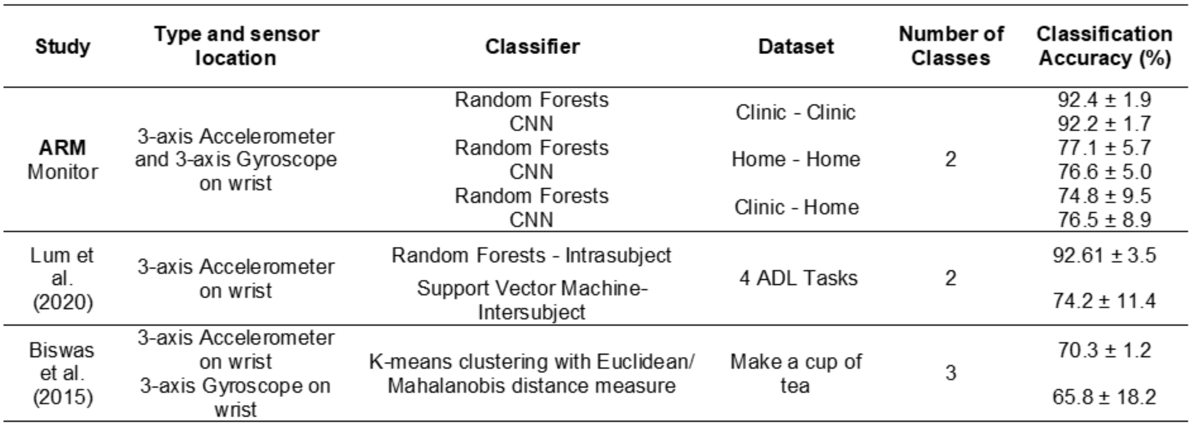
IV. DISCUSSION

In this paper, we presented a prototype for a stroke rehabilitation monitoring system that detects discrete reaching actions in relatively unstructured ADLs in order to facilitate a future application in in-home rehabilitation monitoring of arm function following stroke. We trained two machine learning classifiers, Random Forests and Convolutional Neural Networks, to detect reaching gestures from movement data recorded by a wrist-worn IMU with an accelerometer and gyroscope and characterized the classification performance using two different datasets (functional assessments and ADL) which were designed to emulate two key use environments (clinic and home) for applying such a classifier.

The classifiers were most accurate when trained using discrete tasks from standardized functional assessments and tested on these same tasks (Clinic-Clinic tests). In turn, the Clinic-Home configuration likely had the lowest performance because the classifiers were trained on a dataset that was substantially different from the test dataset. Since the Clinic-Home configuration is most relevant to the proposed implementation of these classifiers in a home monitoring system, it is important to consider whether the performance we obtained here would be sufficient to justify such use.

Existing methods to capture arm activity outside of the clinic, such as the Actical activity counts, fail to distinguish between functional and non-functional arm movements and are therefore insufficient to monitor rehabilitation at home. We have seen further proof of this during our study, since the Actical monitor calculated higher counts during passive (non-functional) arm movements associated with walking.

TABLE IV
COMPARISON OF ARM MONITOR TO LITERATURE



Study	Type and sensor location	Classifier	Dataset	Number of Classes	Classification Accuracy (%)
ARM Monitor	3-axis Accelerometer and 3-axis Gyroscope on wrist	Random Forests CNN Random Forests CNN Random Forests CNN	Clinic - Clinic Home - Home Clinic - Home	2	92.4 ± 1.9 92.2 ± 1.7 77.1 ± 5.7 76.6 ± 5.0 74.8 ± 9.5 76.5 ± 8.9
Lum et al. (2020)	3-axis Accelerometer on wrist	Random Forests - Intrasubject Support Vector Machine- Intersubject	4 ADL Tasks	2	92.61 ± 3.5 74.2 ± 11.4
Biswas et al. (2015)	3-axis Accelerometer on wrist 3-axis Gyroscope on wrist	K-means clustering with Euclidean/ Mahalanobis distance measure	Make a cup of tea	3	70.3 ± 1.2 65.8 ± 18.2

Successful home monitoring requires unobtrusive classifiers that can discriminate functional movements. Table IV above compares our results with studies that also explored recognition of functional arm movements with wrist-worn sensors in unconstrained environments. These other studies all used at most 1 sensor per arm to classify functional arm gestures in stroke participants; however, they used different sensors, classifiers, and trained their classifiers on different datasets.

Biswas *et al.* [14] collected accelerometer and gyroscope data on the wrist to detect 3 functional tasks (reach and retrieve, lift cup to mouth, and pouring action) in 4 stroke patients on an activity of daily living - “making-a-cup-of-tea”. Their work showed greater accuracy with the accelerometer than the gyroscope data and obtained their best results with a k-means clustering algorithm. Even though our datasets are different, our results compare favorably to theirs in terms of accuracy. A key difference between our studies is that they do not evaluate the system performance on unscripted tasks, and it is therefore difficult to determine how well their classifier would recognize tasks in an unconstrained environment.

Lum *et al.* [23] presented machine learning algorithms that can measure the amount of functional movement during unscripted ADL tasks (laundry activities, kitchen activities, shopping activities, and bed making activities). They acquired data through a single accelerometer worn on the wrist to estimate the % of functional arm use during the activities. Their method represents the most clinically relevant study to date. They obtained a measure of functional movement, as opposed to counting any movement, and they did so in relatively unconstrained environments that better model the conditions at the patient’s home. Their intrasubject accuracy was 92.6% and their intersubject accuracy was 74.2%. Their experimental method is closest to the Home-Home experiment presented in our study since they trained and tested on data obtained during the ADL tasks. We obtained the highest accuracy for our own Home-Home experiment with the Random Forests classifier at 77.1%. Our results are therefore comparable to theirs, providing further proof that machine learning models can be used in combination with wrist-worn sensors to calculate functional activity outside of the clinical setting. In their limitations, they stated that annotating the ADL’s for the training dataset can be burdensome and that they would look to evaluate the use of a reduced activity script. The Clinic-Home experiment presented in our study shows that this is possible since we were able to train classifiers on activities derived from functional assessment such as the ARAT and Fugl-Meyer and recognize functional movements in an ADL with adequate accuracy.

V. LIMITATIONS

The datasets used in this study are small, given that they were obtained from 12 participants who required an average of 20 minutes to complete the functional assessment, and 15 minutes to complete the ADL. They are also skewed by the characteristics of the participant population, which is representative of a mildly impaired population. Classifiers trained in this study may therefore not generalize well to a more impaired population. Furthermore, while the pizza-making task required a wide range of natural reaching movements, it may not represent larger ranges of motion or different trunk and reach configurations. In particular, there are activities such as walking that would be much more prevalent in the home setting than what is captured by the ADL dataset. Additionally, the video annotation was performed by a single person without verification from a second party. A future study would benefit from having more people annotating the data to reduce bias and errors.

VI. CONCLUSION

Overall, the results of our study are likely reasonably indicative of the performance that could be expected of wrist-based motion assessment systems for home monitoring of stroke recovery. They demonstrate that it is possible to identify a fundamental action such as reaching in unconstrained environments. Although the performance of the ARM system is worst when attempting to generalize from training data acquired in structured functional assessments to testing in less constrained environments (Clinic-Home), the reach count and reach time results indicate that there is a reasonably tight correlation in individual participants between the predicted values and the reference values (most results lie within a factor of two of the reference values), and the consistency of the slopes in individual participants suggests that it might be possible to identify an individual 'correction factor' that a clinician could apply to a participant's data to adjust it to more accurately reflect the actual level of activity the participant performed. In addition, our classifiers appear to be relatively immune (compared with the existing Actical device) to irrelevant arm motions such as those induced by walking.

Prior to clinical deployment, some key device design issues would need to be addressed. In particular, the sensor would need to be reduced in size and more appropriately packaged, and protocols for data storage and transmission would need to be developed and implemented. Since the current results are based on only 12 participants, additional training data would need to be obtained, and, ideally, the results evaluated in a more realistic home setting over a longer period of study that should include a full day's cycle of activities of daily living.

In summary, we believe that these results show that a single wrist-mounted IMU-based sensor coupled with an appropriate classifier may be of important value in monitoring in the rehabilitation or home setting of functionally important reaching movements in recovering stroke patients, and this work therefore justifies further development of the system.

Acknowledgment

The authors acknowledge facility support from the Centre for Hip Health and Mobility.

References

[1] J. A. Kleim and T. A. Jones, "Principles of experience-dependent neural plasticity: Implications for rehabilitation after brain damage," *J. Speech, Lang., Hearing Res.*, vol. 51, no. 1, pp. S225–S239, 2008, doi: 10.1044/1092-4388(2008/018).

[2] E. Taub, J. E. Crago, and G. Uswatte, "Constraint-induced movement therapy: A new approach to treatment in physical rehabilitation," *Rehabil. Psychol.*, vol. 43, no. 2, pp. 152–170, 1998, doi: 10.1037/0090-5550.43.2.152.

[3] C. E. Lang, J. M. Wagner, D. F. Edwards, and A. W. Dromerick, "Upper extremity use in people with hemiparesis in the first few weeks after stroke," *J. Neurol. Phys. Therapy*, vol. 31, no. 2, pp. 56–63, Jun. 2007, doi: 10.1097/NPT.0b013e31806748bd.

[4] K. Y. Chen and D. R. Bassett, "The technology of accelerometry-based activity monitors: Current and future," *Med. Sci. Sports Exerc.*, vol. 37, no. 11, pp. S490–S500, Nov. 2005.

[5] G. Uswatte, W. L. Foo, H. Olmstead, K. Lopez, A. Holand, and L. B. Simms, "Ambulatory monitoring of arm movement using accelerometry: An objective measure of upper-extremity rehabilitation in persons with chronic stroke," *Arch. Phys. Med. Rehabil.*, vol. 86, no. 7, pp. 501–1498, 2005.

[6] D. Rand and J. J. Eng, "Disparity between functional recovery and daily use of the upper and lower extremities during subacute stroke rehabilitation," *Neurorehabil. Neural Repair*, vol. 26, no. 1, pp. 76–84, Jan. 2012, doi: 10.1177/1545968311408918.

[7] R. R. Bailey, J. W. Klaesner, and C. E. Lang, "Quantifying real-world upper-limb activity in nondisabled adults and adults with chronic stroke," *Neurorehabil. Neural Repair*, vol. 29, no. 10, pp. 969–978, Nov. 2015, doi: 10.1177/1545968315583720.

[8] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 1–33, 2014, doi: 10.1145/2499621.

[9] A. Moschetti, L. Fiorini, D. Esposito, P. Dario, and F. Cavallo, "Recognition of daily gestures with wearable inertial rings and bracelets," *Sensors*, vol. 16, no. 8, p. 1341, Aug. 2016, doi: 10.3390/s16081341.

[10] L. Gao, A. K. Bourke, and J. Nelson, "Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems," *Med. Eng. Phys.*, vol. 36, no. 6, pp. 779–785, Jun. 2014, doi: 10.1016/j.medengphy.2014.02.012.

[11] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances," *Data Mining Knowl. Discovery*, vol. 31, no. 3, pp. 606–660, May 2017.

[12] H. Junker, O. Amft, P. Lukowicz, and G. Tröster, "Gesture spotting with body-worn inertial sensors to detect user activities," *Pattern Recognit.*, vol. 41, no. 6, pp. 2010–2024, Jun. 2008, doi: 10.1016/j.patcog.2007.11.016.

[13] A. Zinnen, K. van Laerhoven, and B. Schiele, "Toward recognition of short and non-repetitive activities from wearable sensors," in *Ambient Intelligence*, (Lecture Notes in Computer Science), B. Schiele *et al.*, Eds. Berlin, Germany: Springer, 2007, pp. 142–158.

[14] D. Biswas *et al.*, "Recognition of elementary arm movements using orientation of a tri-axial accelerometer located near the wrist," *Physiol. Meas.*, vol. 35, no. 9, pp. 1751–1768, Sep. 2014, doi: 10.1088/0967-3334/35/9/1751.

[15] M. Stikic, K. V. Laerhoven, and B. Schiele, "Exploring semi-supervised and active learning for activity recognition," in *Proc. 12th IEEE Int. Symp. Wearable Comput.*, Sep. 2018, pp. 81–88, doi: 10.1109/ISWC.2008.4911590.

[16] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31314–31338, Dec. 2015, doi: 10.3390/s151229858.

[17] A. Parate, M.-C. Chiu, C. Chadowitz, D. Ganesan, and E. Kalogerakis, "RisQ: Recognizing smoking gestures with inertial sensors on a wristband," in *Proc. 12th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2014, pp. 149–161, doi: 10.1145/2594368.2594379.

[18] R. Yao, G. Lin, Q. Shi, and D. Ranasinghe, "Efficient dense labeling of human activity sequences from wearables using fully convolutional networks," 2017, *arXiv:1702.06212*.

[19] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. 24th Int. Conf. Artif. Intell. (IJCAI)*. AAAI Press, 2015, pp. 3995–4001.

[20] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," 2016, *arXiv:1604.08880*.

[21] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," 2016, *arXiv:1611.06455*.

[22] H. Ismail Fawaz *et al.*, "InceptionTime: Finding AlexNet for time series classification," 2019, *arXiv:1909.04939*.

[23] P. S. Lum *et al.*, "Improving accelerometry-based measurement of functional use of the upper extremity after stroke: Machine learning versus counts threshold method," *Neurorehabil. Neural Repair*, vol. 34, no. 12, pp. 1078–1087, Dec. 2020, doi: 10.1177/1545968320962483.

[24] H. Junker, P. Lukowicz, and G. Troster, "Sampling frequency, signal resolution and the accuracy of wearable context recognition systems," in *Proc. 8th Int. Symp. Wearable Comput.*, Oct./Nov. 2004, pp. 176–177, doi: 10.1109/ISWC.2004.38.

[25] J.-H. Lin *et al.*, "Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke," *Phys. Therapy*, vol. 89, no. 8, pp. 840–850, Aug. 2009, doi: 10.2522/ptj.20080285.

[26] G. Kwakkel *et al.*, "Standardized measurement of sensorimotor recovery in stroke trials: Consensus-based core recommendations from the stroke recovery and rehabilitation roundtable," *Int. J. Stroke*, vol. 12, no. 5, pp. 451–461, Jul. 2017.

[27] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci. Inf. Eng., Univ. Nat. Taiwan, Taipei, Taiwan, Tech. Rep., 2003, pp. 1–12.

[28] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Found. Trends Comput. Graph. Vis.*, vol. 7, nos. 2–3, pp. 81–227, Mar. 2011, doi: 10.1561/0600000035.

[29] O. Amft, H. Junker, and G. Troster, "Detection of eating and drinking arm gestures using inertial body-worn sensors," in *Proc. 9th IEEE Int. Symp. Wearable Comput. (ISWC)*, Washington, DC, USA, Oct. 2005, pp. 160–163, doi: 10.1109/ISWC.2005.17.