Article

Measuring Activities of Daily Living in Stroke Patients with Motion Machine Learning Algorithms: A Pilot Study

Pin-Wei Chen [1,2,†], Nathan A. Baune [1,2,†], Igor Zwir [3,4], Jiayu Wang [3], Victoria Swamidass [1] and Alex W.K. Wong [2,3,5,6,*]

1   PlatformSTL, St. Louis, MO 63110, USA; benny.pinwei.chen@gmail.com (P.-W.C.); nabaune@wustl.edu (N.A.B.); victoria@platformstl.com (V.S.)
2   Program in Occupational Therapy, Washington University School of Medicine, St. Louis, MO 63108, USA
3   Department of Psychiatry, Washington University School of Medicine, St. Louis, MO 63110, USA; zwir@wustl.edu (I.Z.); wangjiayu@wustl.edu (J.W.)
4   Department of Computer Science and Artificial Intelligence, University of Granada, 18010 Granada, Spain
5   Department of Neurology, Washington University School of Medicine, St. Louis, MO 63110, USA
6   Center for Rehabilitation Outcomes Research, Shirley Ryan AbilityLab, Chicago, IL 60611, USA
*   Correspondence: alexwkwong@gmail.com
†   These authors are co-first authors.

Abstract: Measuring activities of daily living (ADLs) using wearable technologies may offer higher precision and granularity than the current clinical assessments for patients after stroke. This study aimed to develop and determine the accuracy of detecting different ADLs using machine-learning (ML) algorithms and wearable sensors. Eleven post-stroke patients participated in this pilot study at an ADL Simulation Lab across two study visits. We collected blocks of repeated activity ("atomic" activity) performance data to train our ML algorithms during one visit. We evaluated our ML algorithms using independent semi-naturalistic activity data collected at a separate session. We tested Decision Tree, Random Forest, Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost) for model development. XGBoost was the best classification model. We achieved 82% accuracy based on ten ADL tasks. With a model including seven tasks, accuracy improved to 90%. ADL tasks included chopping food, vacuuming, sweeping, spreading jam or butter, folding laundry, eating, brushing teeth, taking off/putting on a shirt, wiping a cupboard, and buttoning a shirt. Results provide preliminary evidence that ADL functioning can be predicted with adequate accuracy using wearable sensors and ML. The use of external validation (independent training and testing data sets) and semi-naturalistic testing data is a major strength of the study and a step closer to the long-term goal of ADL monitoring in real-world settings. Further investigation is needed to improve the ADL prediction accuracy, increase the number of tasks monitored, and test the model outside of a laboratory setting.

1. Introduction

Stroke is a leading cause of long-term disability, with nearly 800,000 adults in the U.S. experiencing a stroke annually [1]. Impairments in sensorimotor function, as are common following stroke, negatively impact performance in activities of daily living (ADLs) [2]. Almost half of the stroke survivors experience limitations in ADLs [3]. These limitations are a key concern among clinicians and patients. Research has shown strong relationships between performance in ADLs and patients' quality of life [4] and the risk of re-hospitalization [5,6] after stroke. ADLs are complex sensorimotor activities involving dynamic spatial-temporal coordination of our limbs and trunk. Developing effective tools for monitoring ADLs and complex bodily movement following stroke could provide a wealth of clinically relevant information useful for tailoring therapy post-stroke.

However, clinical assessment of ADLs in stroke is limited to self-reported or clinician-rated scales, such as the Functional Independence Measure or the Barthel Index. While these measures have been clinically validated, they are mostly retrospective and susceptible to reporting bias and error [7–9]. An earlier study involved elderly patients and their families or nurses completing the Lawton Personal Self Maintenance scale and the Instrumental ADL (IADL) scale. They found that patients consistently rated themselves less disabled on both measures than other raters [10]. Another study found similar results in which patients reported less disability in ADL scales than family raters. The trend was held when comparing patient self-reports to researcher observations [11]. These findings indicated that patients might have trouble noticing their disability or underplay their disability. Furthermore, because these measures are rated on an ordinal scale, they may lack the resolution to detect subtle changes, limiting the ability to monitor a patient's recovery precisely. Due to patient burden, costs, or contextual factors, clinicians infrequently implement clinical measures, so this approach is inadequate for characterizing the day-to-day variability of an individual's function [12].

Fortunately, the last decade has seen massive growth in the capacity to collect an array of physiologic data for monitoring the health and functioning of an individual via sensors, such as those commonly found in smartphones or wearable devices [13]. Although wearable technologies have enabled automatic and continuous measurement of metrics useful for predictive medicine, such as energy expenditure or pedometers, current utilization of wearable data provides an incomplete depiction of an individual's performance of ADLs, especially for aging and disabled populations. Mobility is often limited in these populations, reducing the utility of step-counters, heart rate measures, and other crude metrics to accurately measure an individual's capacity to participate in meaningful ADLs. Moreover, metrics derived from wearable sensors (e.g., acceleration or angular velocity) do not have direct clinical meaning and are, therefore, difficult to use for clinical decision making. It is becoming imperative to develop better technology that provides an objective and clinically-valued measurement of an individual's ADLs.

Human activity recognition has garnered intense research interest over the past few decades and can be generally separated into inertia-based detection and video-based detection approaches. On the one hand, video-based activity detection has been growing tremendously and has shown promising results (please refer to these reviews for details [14–16]). However, video-based activity detection raises privacy concerns due to potential breach of data or data misuse [17]. On the other hand, inertia-based activity recognition is less intrusive, and has shown significant improvements over time [17]. Inertia-based activity recognition systems utilize IMU's (inertial measurement units consisting of an accelerometer and gyroscope) to measure body kinematics directly. These direct measures may provide additional information for recognizing post-stroke activities. This technology would allow researchers and clinicians to monitor rehabilitation progress, validate treatment efficacy, and detect functional decline early. Such radical change in measurement may further heighten our ability to track real-world outcomes in rehabilitation and provide accurate measures for decentralized (fully remote) clinical trials in the future.

Although the inertia-based activity recognition approach has received increasing attention [18–21], prior research was based on a large repository database of able-bodied individuals. In addition, previous research did not directly record mobility-impaired individuals (e.g., individuals who experienced a stroke) under a clinical setting for machine learning (ML) purposes. In addition, most studies have performed internal validation, where a portion of the same data set is used to test a model created from the remaining data. These are barriers to the eventual application of ADL monitoring in real-world settings. Therefore, this study aimed to develop a novel prediction model based on ML algorithms and to determine the accuracy of detecting different ADLs performed by stroke survivors using wearable sensors. We conducted this study in a simulation living room and kitchen. Lastly, we collected independent training and testing data to perform external validation, which more closely imitates real-world prediction conditions.

2. Materials and Methods

2.1. Participants

Participants were community-dwelling adults with stroke. We recruited 11 stroke survivors from the stroke registry at Washington University School of Medicine, a database of individuals who consented to future research participation at the time of their stroke hospitalization. Inclusion criteria included: (1) age 18+; (2) English-speaking; and (3) mild stroke as defined by baseline National Institutes of Health Stroke Scale (NIHSS) score from 0 to 5. We chose to study patients with mild stroke because motor function and ability to accomplish basic ADLs are only minimally affected, increasing the chance that the protocol will be fully completed in an appropriate time and allowing us to adhere to the project budget. Exclusion criteria: (1) previous neurologic or neuropsychiatric disorder (e.g., dementia, schizophrenia) that makes interpretation of the self-rated scales difficult; (2) Short Blessed Test score >8 (indicating significant cognitive impairment); (3) history of moderate disability prior to stroke (pre-morbid Rankin Scale score <3); (4) vision that is poorer than 20/100 (as determined by the Lighthouse Near Visual Acuity Test); (5) Apraxia screen of Tulia <9; and (6) evidence of severe aphasia (NIHSS aphasia item >2). Eligible participants participated in our study at an ADL simulation lab in the Program in Occupational Therapy across two study visits. All participants were at least six months post their stroke incident, meaning their functional recovery had largely plateaued.

2.2. Procedures

During each visit, we fitted participants with five inertial measurement units (IMUs) to collect accelerometer and gyroscope data: one on each wrist, one on each of their upper arms, and one on their hip. The IMU used in this study were Apple Watch Series 3 (Apple Inc., Cupertino, CA, USA). We developed the system using the Apple Watches due to their affordability, commercial availability, and ability to transfer motion data wirelessly over users WiFi network. We utilized two study visits to collect independent training (visit two) and testing (visit one) data sets. During visit one, we used a method similar to that introduced by Bao and Intille [18] for capturing naturalistic behaviors. Participants engaged in a series of ADL tasks following a standardized script for the examiner to provide minimal guidance. The examiner informed participants what tasks they would perform next but did not comment on how to achieve them. For example, the examiner would ask participants to cook stir-fry following a recipe (which would require gathering ingredients, chopping, cooking), then to eat a serving (which would require plating the food, retrieving eating utensils, and sitting down to eat). Participants were allowed to shift between activities naturally and perform actions like chopping and cooking as they preferred. The experimenter aimed to provide as little guidance as possible, giving cues for the next steps and answering questions; otherwise, participants operated primarily independently. This care was taken to most closely mimic real-world scenarios of ADL prediction. Participants performed these semi-naturalistic activities to record testing data for evaluating the ML algorithms. We used surveillance cameras to record and label all activities from visit one and later review for possible errors.

During visit two, participants performed four minutes of each atomic activity (a total of 19 atomic activities; Table 1). An atomic activity is a simple movement involved in ADL tasks (e.g., stirring a pan or chopping vegetables are two atomic activities involved in cooking). Atomic activities require participants to continuously perform specific movements with experimenter guidance, which contrasts with semi-naturalistic activities where we provided participants with a high-level goal (e.g., cook pancakes) but did not give explicit performance instructions. Participants performed atomic activities repeatedly for four minutes; for example, the participant would stir pancake batter for four minutes, place/retrieve spices on/from a shelf for four minutes, or fold laundry for four minutes. We collected atomic activity performance data to train our ML algorithms. The experimenter recorded the timing of atomic activity performance using in-house software (henceforth, activity labeling software) on a seventh-generation iPad. While four minutes of atomic

activity data were collected per activity, participants could perform activities in multiple blocks to reduce boredom or fatigue due to the homogenous repeated motion. Participants performed the semi-naturalistic activities first (visit one) before they completed the atomic activities (visit two) to reduce possible bias. Specifically, we avoided participants performing atomic activities first because the homogenous repeated motion might influence the way participants performed the semi-naturalistic activities.

Table 1. Atomic activities with categories label between activities of daily living (ADL) or instrumental activities of daily living (IADL).

Atomic Activities	ADL/IADL
Brush Teeth	ADL
Mixing Powders	IADL
Spreading Butter/Jam	IADL
Eating with Hands	ADL
Eating with Utensils	ADL
Buttoning shirt/coat	ADL
Put and take off the coat	ADL
Moving items horizontally	ADL
Reaching Up	ADL
Reaching Down	ADL
Washing Dishes	IADL
Chopping	IADL
Pan Stirring	IADL
Serve on a plate	IADL
Sweeping	IADL
Vacuuming	IADL
Wiping horizontal surface	IADL
Wiping vertical surface	IADL
Folding clothes	IADL

Prior to this study, we tested on healthy adults and achieved adequate accuracy when collecting only two minutes of atomic activity data per activity. Nevertheless, we collected four minutes per atomic activity in this study to increase the amount of training data. We anticipated this amount of data to be beneficial due to potential variability between post-stroke participants. We could ensure adequate data remained for each activity in any case that segments of data had to be removed due to errors (e.g., technical difficulties with IMU recording). The four-minute was chosen as the upper-limit as a precaution to avoid participant over-exertion. We could further ensure data were collected for all activities within the allotted three-hour study sessions. The study sessions were all facilitated by a licensed occupational therapist trained to monitor these activities and ensure patient safety.

We chose the 19 activities (Table 1) based on several criteria. We determined the final ADL and IADL list through consultation with therapists and physicians specialized in stroke. Activities chosen were: (1) commonly used to assess post-stroke functional activity, (2) amenable to performance in our simulated living environment, and (3) feasible for mild stroke patients to perform during the allotted three-hour testing sessions. This study received ethics approval from the institutional review board at Washington University. All participants provided written consent and received an honorarium to acknowledge their research contribution.

2.3. Data Analysis

We used descriptive statistics to characterize the study sample. We combined initial data from five devices for each participant and assigned activity labels.

These predictor variables included the translational and rotational acceleration along three axes (x, y, z) across each of the five sensors, resulting in 30 variables (Table A1 of Appendix A). We used three-second epochs to summarize the raw data into feature space with time-domain (e.g., mean, standard deviation, autocorrelation, and slope) and frequency-domain features (see Figure A1 of Appendix A for one variable as an example). This technique provided stationarity for the time series, allowing each activity to be analyzed as a stationary stochastic process in model development [22]. Data were preprocessed, including slicing data with an optimal window size to avoid redundancies, input variables were normalized and then compressed into 6012 data points. We chose three-second epochs based on our previous testing in healthy adults, where three-second was the optimal epoch length while testing five epoch lengths (integers 1 through 5).

Motion data from the semi-naturalistic activity session (visit 1) was labeled using the data from our activity labeling software. We used data from visit 2 (atomic activities) for the training, whereas data from visit 1 (semi-naturalistic activities) for the testing. We implemented t-Distributed Stochastic Neighbor Embedding (t-SNE) analysis and cluster analysis to analyze the samples' distribution in terms of the first and second visits.

The classes/tasks' sizes for activities substantially vary among the observed subjects, so we rebalanced the training data sets to prevent over-teaching the classifier only to predict the major (negative) classes. We accomplished this using both the Synthetic Minority Oversampling Technique (SMOTE), a bootstrapping algorithm that provides more data points for the smaller class based on variable distributions of that class [23–25] and the Extended Nearest Neighbor (ENN) algorithm that downs size the larger class [25] for the training data. It should be noted that both SMOTE and ENN were only used on the training data after the data were split into the training and validation data sets. Moreover, since learning the hyperparameters involved the 70–30% hold-out validation, SMOTE and ENN was used only in 70% hold-out. We also used the Shapley Additive exPlanations (SHAP) to identify the contribution of the individual IMU features in the model predictions [26,27].

2.4. Classified Model Development

We achieved the primary classification model with the gradient boosting model (GBM), XGBoost [28]. This method has been shown to have improved supervised classifier model formalization, compared to other GBMs or random forest algorithms, and better efficiency and controls for overfitting [29].

Hyperparameters supported by the XGBoost package were fine-tuned using the Tree Parzen Estimator (TPE) Bayesian optimization algorithm [30] to find the best combination of parameters based on the 70–30% hold-out validation. The fitness loss function was defined by the AUC [29]. Data were randomly partitioned into two sets based on the hold-out sampling partition used for the internal training validation test with unbalanced classes of atomic activities.

There was no need for other sampling methods rather than hold-out because the external semi-naturalistic sample was completely disjoint, including variability in terms of times of tasks, order, completeness, and the presence of the training tasks. The external independent validation dataset based on semi-naturalistic activities was not seen at any training session or data processing. Tasks were selected by following a standard stepwise regression using XGBoost, where models were fitted based on the choice of predictive variables carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some pre-specified criterion. The stop iteration criteria was a performance of 90% $\pm$ 2%.

All scripts were written in python, and the XGBoost package and the other classifiers were used with a python sklearn API (https://scikit-learn.org (accessed on 9 December 2020)) [28,31]. For parameter fine-tuning, we used the hyperopt package (https://scikit-

learn.org/stable/modules/grid_search.html (accessed on 9 December 2020)). We used the imbalance package to apply SMOTE and ENN. We used Sklearn for all data partitions and the calculation of the metrics.

We selected the final set of features, as estimated above, and then tested multiple ML models: Decision Tree, Random Forest, SVM, and XGBoost to determine the optimal classification model. Default parameters for each of these methods were learned as described above (sklearn-scikit) by optimizing the AUC metric. We assessed model performance using accuracy, recall (sensitivity), precision (positive predictive value), and ROC (receiver operating characteristic) curve (AUC) metrics [32]. The loss function was based on AUC.

3. Results

3.1. Characteristics of Study Participants

Table 2 provides an overview of participant demographic characteristics. The majority of participants were males (72%), right-handed (91%), and experienced an ischemic stroke (100%). The average age was 60 years old. 55% (*n* = 6) participants experienced a right hemispheric stroke, 36% (*n* = 4) experienced a left hemisphere stroke, and one participant did not have information on the side of stroke. The average time since the stroke incident was 2.76 years (SD = 1.73).

Table 2. Demographic characteristics of the participants (*n* = 11).

Characteristics	Mean (SD)/Count
Age	59.64 (9.04)
Years since stroke	2.76 (1.73)
NIHSS [1] Total Score	1.45 (1.29)
Sex	8 males; 3 females
Race	6 Caucasian; 4 African American; 1 unknown
Education	4 high school graduates; 5 some college; 2 bachelor's degree
Handedness	10 right; 1 left
Stroke Type	11 ischemic; 0 hemorrhagic
Stroke Side	6 right-hemispheric; 4 left-hemispheric; 1 unknown

[1] NIHSS = National Institutes of Health Stroke Scale.

3.2. Data Extraction and Model Development

The total duration of the observed data was 153,159 s. We divided data into three-second epochs; thus, the total epochs were 51,053. It should be noted that we encountered a hardware malfunction for one participant's visit, but we did not identify this malfunction until the experiment was completed. As a result, we excluded a vast amount of data of this participant for subsequent analyses. For feature selection, we first utilized the t-SNE graph to understand whether there were dissociations between training data from visit two and testing data from visit one (Figure 1). We found no significant separations between the two data segmentation, suggesting no dissociations between the two datasets. This result indicates that we can safely use the training set based on atomic measurements in visit 2 for predicting semi-naturalistic activities in visit 1.
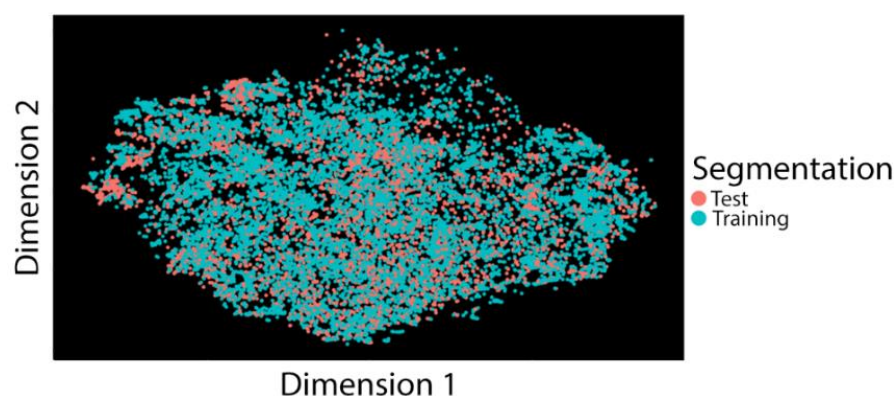
Int. J. Environ. Res. Public Health **2021**, *18*, 1634

7 of 16

Figure 1. t-Distributed Stochastic Neighbor Embedding (t-SNE) graph showing segmentation for both training and test data sets.
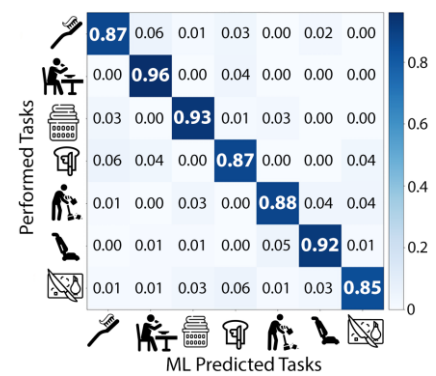
Figure 2 shows three confusion matrices corresponding to classifications based on 19, 10, and 7 tasks. We selected tasks retained in the model based on the stepwise regression process, in which we removed one ADL/IADL task on each run with the lowest accuracy. Using the XGBoost algorithm, we achieved an average accuracy of 97% on the training set and an accuracy of 90% in an independent test based on seven tasks in a sample composed of all subjects while performing semi-naturalistic activities. We also obtained 0.91 for recall (sensitivity), 0.83 for precision, and 0.98 for AUC in the independent test set. The seven tasks include cutting, vacuuming, sweeping, spreading jam or butter, folding laundry, eating, and brushing teeth. Ten tasks produced 86% accuracy on the training set and an 82% accuracy on the independent test set. Ten tasks include the previous seven tasks plus: Taking off/putting on a shirt, wiping cupboards, and buttoning a shirt. We further computed the t-SNE distributions of the ten tasks and found many of them clustered together (Figure 3), suggesting that these are predictable structured classes. SHAP indicated that, among the five IMUs, both wrists and the hip captured the most critical information for activity recognition. The important features contained gyroscopic standard deviation and accelerometer standard deviation and mean. Table 3 shows the performance evaluation metrics across ML models. We found that XGBoost optimizes or equates with methods that produce the best results. These results provided initial evidence that ADL/IADL functioning can be predicted with adequate accuracy by using wearable sensors (IMUs) and ML algorithms. We also used the SHAP values to identify the contribution of the individual IMU features to predict activities in the model prediction (Figure 4). Among 5 IMUs, both wrists and the hip, capturing the most critical information for predicting activities. Essential IMU features included standard deviation and mean features from gyroscope and accelerometer.

Int. J. Environ. Res. Public Health **2021**, *18*, 1634

8 of 16
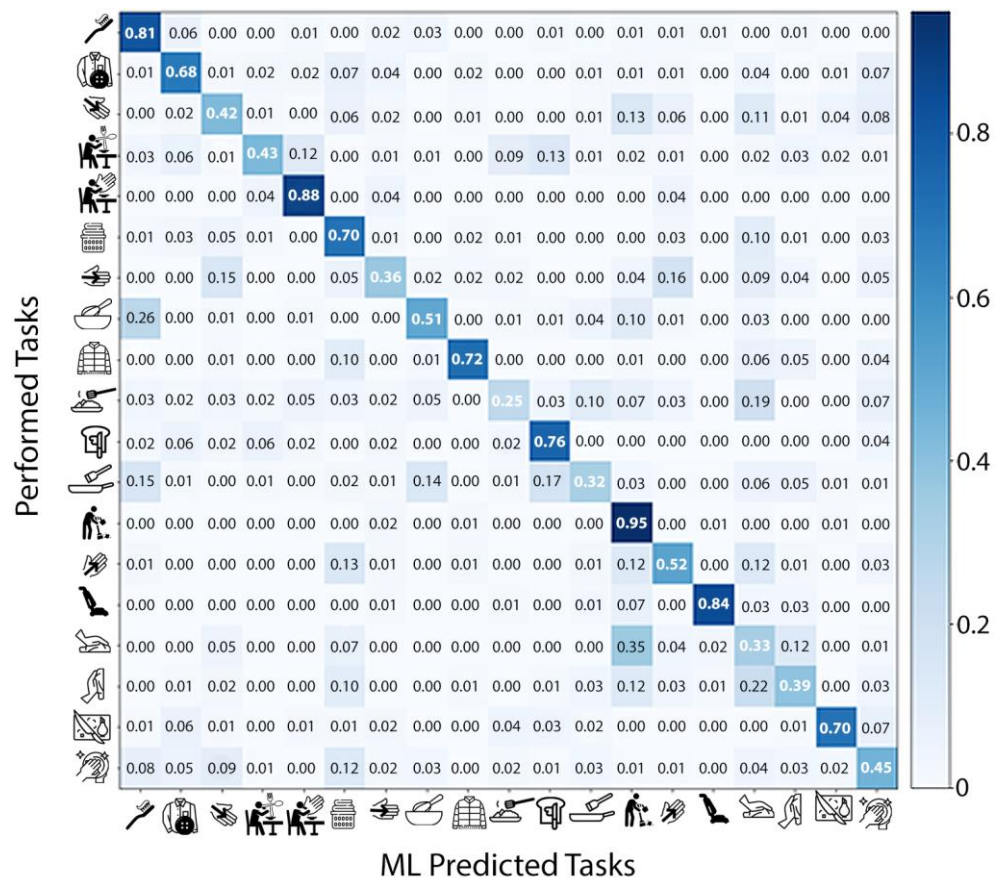
(**a**)
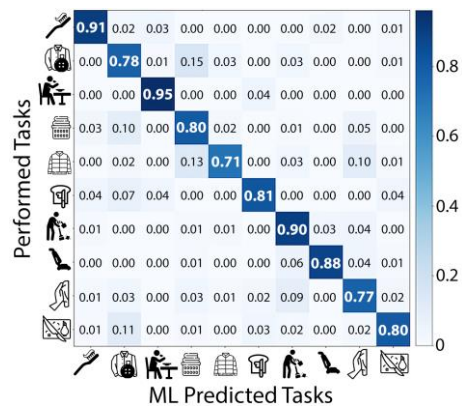
(**b**)

(**c**)

Figure 2. Confusion matrices corresponding to classifications based on (**a**) 19 tasks, (**b**) 10 tasks, and (**c**) 7 tasks. Tasks selection was performed following a standard stepwise regression process using XGBoost.
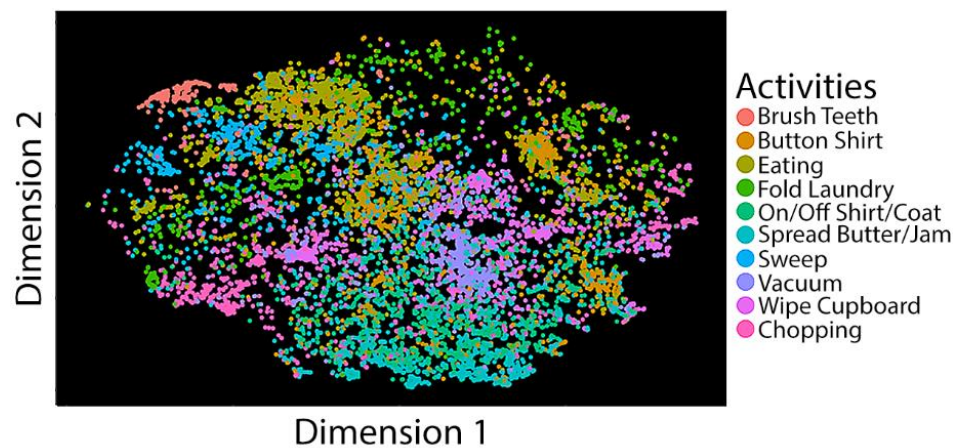
Figure 3. t-SNE graph for the ten ADL tasks. Dots indicate data from different ADL tasks.

Table 3. Performance metrics of seven ADL tasks classification across machine learning (ML) models.

Performance Metric [1]	Decision Tree	Random Forest	SVM	XGBoost
Training Set				
Accuracy	0.56	0.79	0.97	0.97
AUC	0.74	0.88	0.99	0.98
Precision	0.50	0.80	0.97	0.97
Recall	0.56	0.79	0.97	0.97
Test Set				
Accuracy	0.43	0.80	**0.90**	**0.90**
AUC	0.68	0.89	0.95	**0.98**
Precision	0.47	0.84	**0.92**	0.83
Recall	0.43	0.80	0.90	**0.91**

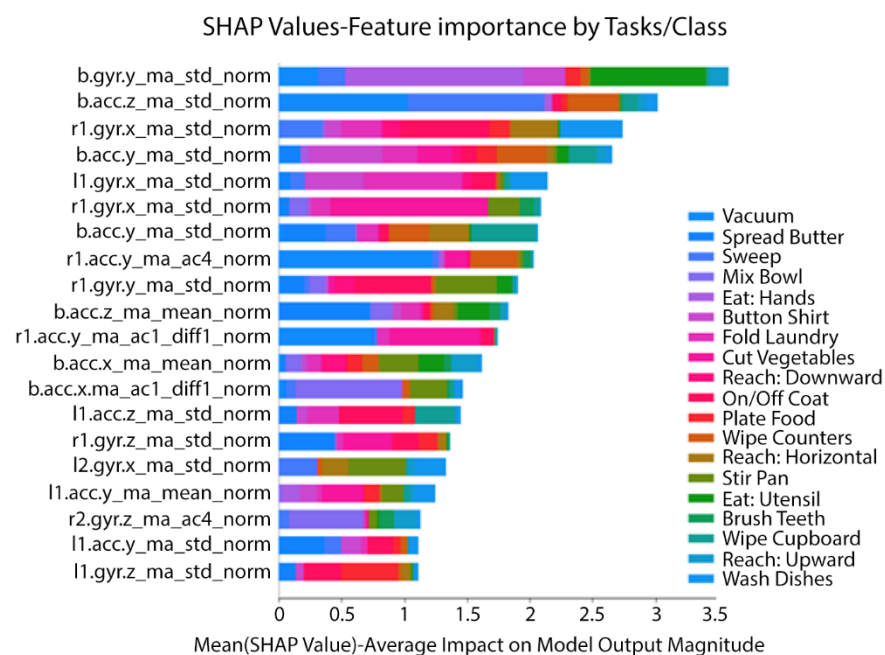[1] Note: Numbers in bold represent the best values across ML models.

Figure 4. Influence of various inertial measurement units (IMU) features for predicting ADL tasks on the XGBoost model.

4. Discussion

Using motion-based data collected from IMUs, we have developed and validated ML algorithms to recognize a list of ADLs among a sample of community-dwelling stroke survivors. A significant contribution of our study is the use of independent training and testing data sets. We asked participants to complete semi-naturalistic activities in an obstacle course design (e.g., unload groceries and then prepare and eat stir-fry using ingredients). We also asked them to complete atomic activities recorded within a block design (e.g., four minutes of chopping vegetables). After that, we trained the ML algorithms with the atomic activities and tested them with semi-naturalistic activities. We found that our ML algorithms could recognize a set of activities in a semi-naturalistic environment with adequate accuracy. Enabling participation in daily activities is the ultimate outcome in rehabilitation. These results provide initial evidence that motion-based sensors and ML, especially the XGBoost and SVM algorithms, can predict post-stroke daily activities.

Among all 19 activities included in the ML model, we found the top three activities that can be accurately classified are sweeping (0.95), eating with hands (0.88), and vacuuming (0.84), whereas other activities that had the lowest accuracy are serving on a plate (0.25) and pan stirring (0.32). We have inspected the raw data and found that these high accuracy activities often involve highly repetitive movements that can be tracked with frequency-domain features. In contrast, the low accuracy activities consist of high variations in the data (i.e., movements are less repetitive or occasionally occur within a 3 s epoch). For example, serving on a plate is an activity that often involves multiple movements that happen in a short time and the motion sequence is not highly repetitive. Participants would grab the spatula, raise the food item, and then place the food item on a plate occurring approximately less than 10 s. The motion sequence of these steps is not repetitive during the semi-naturalistic recordings. Pan stirring is an activity in which participants would stir the food once and stop for a while before the next stirring, making the model hard to distinguish these motions within the 3 s epoch window.

Human activity recognition is a growing field. A prior review summarized all accelerometer-based human activity recognition research and found that past research focuses primarily on mobility activities (e.g., sitting and standing) or postures [33]. Another review paper summarized all applications of activity recognition systems. It found that ML algorithms utilizing video recordings (e.g., Infra-red and RGB-D) have tremendous improvement in recognizing the types of activities [17]. Further review of video-based detection systems can be found in these papers, summarizing the current status of the research [14–16]. Although video-based systems are useful for security or surveillance solutions, it is not as valuable when utilizing it for patient monitoring due to privacy concerns.

Human activity recognition using IMUs or other wearable sensors has remained a likely alternative to video-based systems. Most wearable sensors allow measurement of an individual's physiological signals (e.g., heart rate) and direct kinematics (e.g., intensity and moderate-to-vigorous physical activity). These measures may provide additional information for recognizing activities among older adults and individuals with physical disabilities, such as those after the stroke. However, there are not many accelerometry-based ML research that directly records mobility impaired individuals. A study involving a group of stroke survivors utilized wearable smartphones to predict six "activity" labels: Sit, stand, lie, stairs, large movements (including walking, small steps, and opening doors), and small movements. This research was able to achieve high accuracy in predicting these "activities" [34]. Another study evaluated the likelihood of recognizing postures of individuals with a stroke while performing activities using smartphones. This research found that increasing the complexity of activities decreased model accuracy in detecting postures [35]. These two studies have demonstrated the potential of using mobile technologies and ML methods for human activity recognition.

Many of the previous studies utilized a database from able-bodied individuals [19–21]. Models built using data from able-bodied individuals may not generalize to stroke survivors. Indeed, models built solely from mild stroke patient data may not generalize to

moderate or severe stroke. The generalizability of models is a major concern in building clinically valid models. The current study was designed as a pilot study to present the feasibility of our approach to ADL/IADL monitoring. We restricted this study to mild stroke patients to ensure they could perform all activities and in the allotted study session times. We acknowledge such limitations and in future studies we will need to address these concerns. For example, it may be possible that a single model across mild, moderate, and severe stroke survivors is feasible or perhaps different groups require unique models. An additional common shortcoming in prior research is the reliance on internal validation to assess the accuracy of models. An example of internal validation includes building a model using one portion of the data set and testing it with the remaining portion. This approach is likely to reduce the ecological and clinical relevance of a model. Using clean data (no interruptions or spurious data) to train the model is important, through predicting activity using clean data may also exaggerate the model utility. Naturalistic performance is not clean. For example, individuals often transition between activities frequently, pausing to talk and communicating with their hands while doing so. A major contribution of our study is the use of clean data to train our models and an independent semi-naturalistic dataset to validate the models. The goal was to match real-world performance as closely as possible for our testing dataset. We acknowledge that the simulated living environment and the structured script to guide the activities are not a perfect model of natural activity. However, we took many precautions to better match natural performance conditions; the high degree of accuracy we achieved using data from stroke patients and through external validation using an independent dataset are major strengths of our study. In future studies we plan to transition into performance in the user home to match real-world performance even more closely.

Along with other activity recognition studies, our current study added to the literature by harnessing motion sensors and building ML algorithms to detect different types of at-home activities and directly collecting activity data in stroke survivors. Our research provides initial evidence to predict (i.e., recognize) ADL functioning in stroke. More importantly, the methodology and technology developed to collect and analyze data in this study may translate into inpatient rehabilitation facilities for other populations because these facilities often have ADL simulation rooms.

During the experiment, we did not observe any particular activity that participants had significant difficulty completing. This is likely due to our participants, who are mild stroke survivors with minimal motor and ADL problems. Nevertheless, we observed that some participants demonstrated slight difficulties with certain bilateral activities. We reviewed the raw data and found certain activities have a higher variation of inertial intensity. Specifically, spreading butter/jam and eating with hands have the highest variations of inertial intensity, suggesting that these two tasks demand more irregular motions performed by the upper limbs. On the other hand, wiping the cupboard or pan stirring has a lower variation, suggesting these tasks require less fine-motor ability. Although differentiation of difficulty level of various activities is not the scope of our current study, our current ML model could classify some of them with high accuracy.

Furthermore, to explore our technology's clinical utility in characterizing patients' ADL performance, we have conducted a supplementary analysis of two selected activities that typically require bilateral use of the upper limbs (Figure A2 of Appendix B). This analysis aimed to explore the degree of usage of upper limbs while performing specific ADL tasks. According to prior research [36,37], bilateral limb usage impairment is common for individuals after a stroke. To quantify the real-world activity, an accelerometry-based outcome can be used to examine each upper limb's intensity of usage while a patient performs a bilateral task [38]. This outcome measure further helps clinicians and researchers quantify the difficulty level of various activities. For example, clinicians may use the plots seen in Figure A2 to understand the dominant and non-dominant limbs' usage demand across multiple activities. As seen, when compared to the "spreading butter/jam, "chopping vegetables" demanded higher intensity on the dominant hand relative to the

non-dominant hand. Knowing the usage demand of various activities can help develop the treatment plan by selecting appropriate ADL tasks for interventions that match the patient's motor function. Other possible outcomes are the amount of time spent on specific activities and the repetition and variation among those activities.

There are several limitations to this study. First, we have a small sample size for model building. As this is a pilot study, the current research provides the first step in designing a better paradigm for large-scale data collection. We suspect that it is essential to collect high-quality movement data from the target population. Thus, we will continue to hone our data collection techniques. For example, we will further improve our customized activity labeling software, which allows us to label movement data into activity categories and reduce the number of resources required to conduct ML research, a major barrier in activity recognition research. Second, we collected activity data in a semi-naturalistic environment instead of participants' homes. This methodology may reduce the generalizability of detecting activities in real-world settings. Nevertheless, this study was designed to test the data collection method and provide initial proof-of-concept to detect daily activities from individuals with a stroke. The semi-naturalistic environment serves as an intermediate step to gain knowledge and experience to record data before testing it in participants' homes. Further investigation is also needed to improve the ADL prediction accuracy, increase the number of tasks monitored, and test the model outside of a laboratory setting.

Developing more capable models and verifying their efficacy in a larger and more diverse sample of stroke patients will be a critical next step, although these initial results are highly promising. We will work closely with clinicians and patients as we move forward. We will collaborate to determine the most appropriate activities to monitor and the most valuable ways to present the findings. We understand this user-centered approach will be of utmost importance if we hope for our technology to solve some of the most critical issues plaguing rehabilitation practice and research. If successful, the advancements could benefit stroke survivors and a wide range of individuals with disabilities. Remote monitoring of activities crucial to functional independence may bring clinicians and researchers much closer to talent the challenging goal of individualized/personalized medicine.

5. Conclusions

Enabling patients' daily activity and participation in life is the ultimate goal for rehabilitation. Current results in this pilot study have shown the possibility to monitor task-specific ADLs from IMU data collected in a simulated setting and modeled with the ML methodology. With the ADL data reflecting patients' real-world functional performance, clinicians and researchers can have insight into whether the treatment positively impacts an individual's everyday life.

Appendix A

Table A1. A total of 30 predictor variables including the translational and rotational acceleration along three axes (x, y, z) across each of the five sensors.

| Variables	Variable Names	Descriptions
1	l1.acc.x	Left Wrist, Accelerometer, X-axis
2	l1.acc.y	Left Wrist, Accelerometer, Y-axis
3	l1.acc.z	Left Wrist, Accelerometer, Z-axis
4	l1.gyr.x	Left Wrist, Gyroscope, X-axis
5	l1.gyr.y	Left Wrist, Gyroscope, Y-axis
6	l1.gyr.z	Left Wrist, Gyroscope, Z-axis
7	l2.acc.x	Left Arm, Accelerometer, X-axis
8	l2.acc.y	Left Arm, Accelerometer, Y-axis
9	l2.acc.z	Left Arm, Accelerometer, Z-axis
10	l2.gyr.x	Left Arm, Gyroscope, X-axis
11	l2.gyr.y	Left Arm, Gyroscope, Y-axis
12	l2.gyr.z	Left Arm, Gyroscope, Z-axis
13	r1.acc.x	Right Wrist, Accelerometer, X-axis
14	r1.acc.y	Right Wrist, Accelerometer, Y-axis
15	r1.acc.z	Right Wrist, Accelerometer, Z-axis
16	r1.gyr.x	Right Wrist, Gyroscope, X-axis
17	r1.gyr.y	Right Wrist, Gyroscope, Y-axis
18	r1.gyr.z	Right Wrist, Gyroscope, Z-axis
19	r2.acc.x	Right Arm, Accelerometer, X-axis
20	r2.acc.y	Right Arm, Accelerometer, Y-axis
21	r2.acc.z	Right Arm, Accelerometer, Z-axis
22	r2.gyr.x	Right Arm, Gyroscope, X-axis
23	r2.gyr.y	Right Arm, Gyroscope, Y-axis
24	r2.gyr.z	Right Arm, Gyroscope, Z-axis
25	b.acc.x	Hip, Accelerometer, X-axis
26	b.acc.y	Hip, Accelerometer, Y-axis
27	b.acc.z	Hip, Accelerometer, Z-axis
28	b.gyr.x	Hip, Gyroscope, X-axis
29	b.gyr.y	Hip, Gyroscope, Y-axis
30	b.gyr.z	Hip, Gyroscope, Z-axis
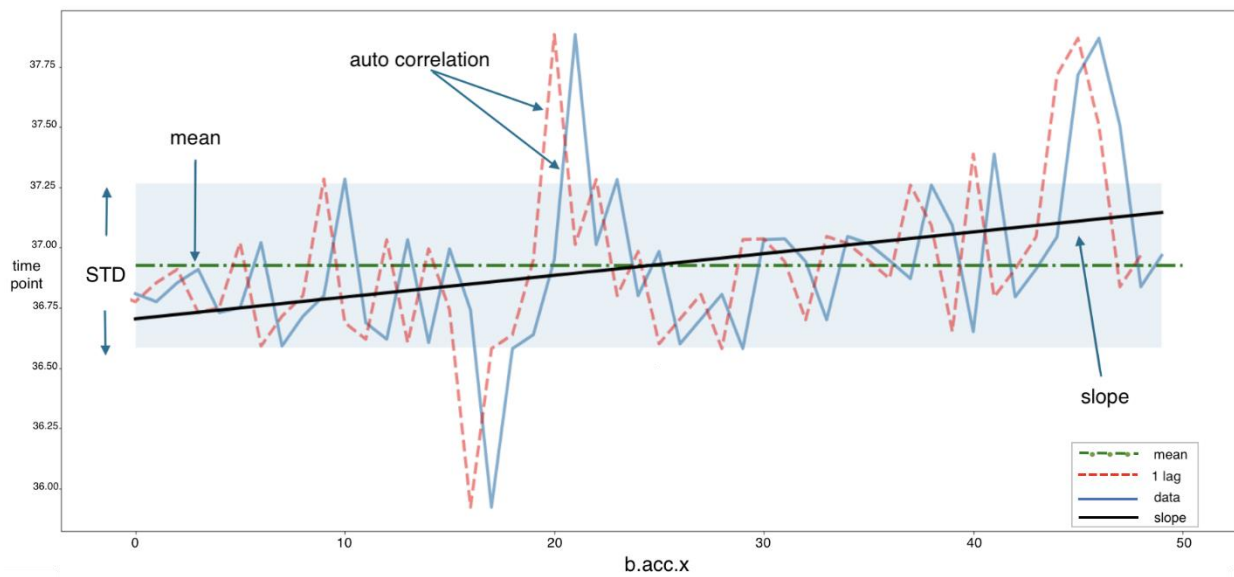
Figure A1. Four statistical features were built, based on continuous variables that accounts for stationarity in temporal processes. The mean represents the overall mean of a subject's entire task. The standard deviation (std) corresponds to the variance of a subject's behavior throughout the entire task. The auto correlation is the correlation between a feature and its corresponding one time-point (5 min) lagged data (1 lag). The slope is the trend captured by a linear regression throughout the entire stay.
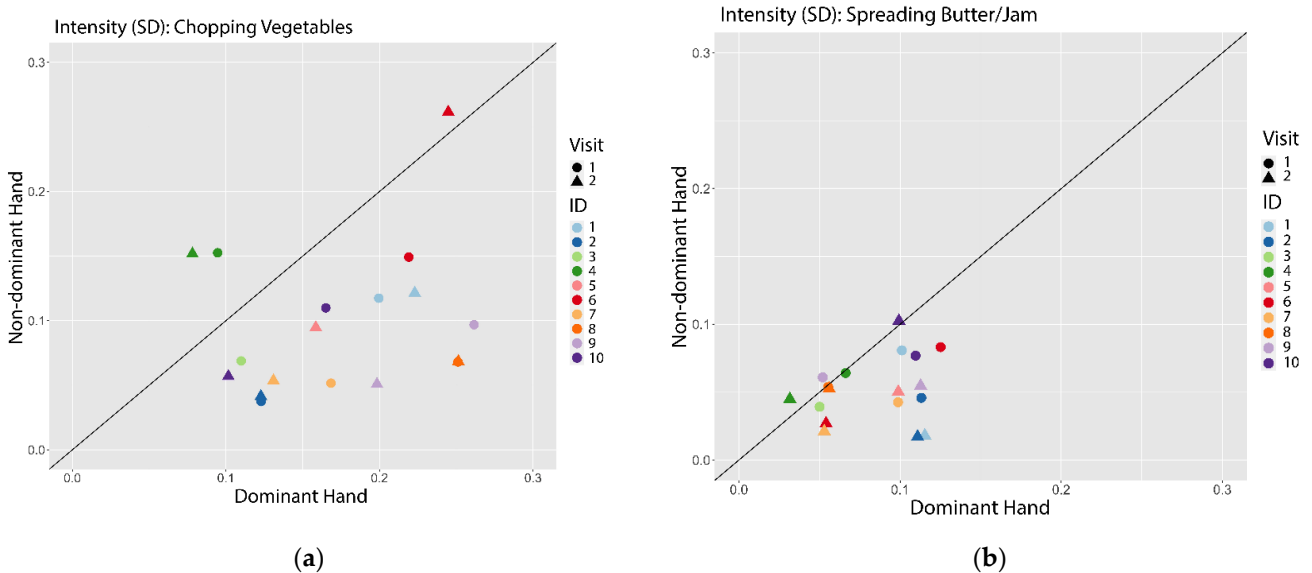
Appendix B

(**a**)　　　　　　　　　　　　　　　　　　　　　　(**b**)

Figure A2. Performance of bilateral ADL tasks: (**a**) Chopping vegetables; (**b**) Spreading butter. Colors indicate different participants. Shapes indicate different visit dates. The black line is the equal standard deviation of the intensity between the non-dominant and dominant hands.

References

1.　Virani, S.S.; Alonso, A.; Benjamin, E.J.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Delling, F.N.; et al. Heart Disease and Stroke Statistics—2020 Update: A Report from the American Heart Association. *Circulation* **2020**, *141*, e139–e596. [CrossRef]

2.　Shamay, N.S.; William, T.W.; Patrick, K.W.; Philip, T.T.; Jefferry, W.C. Sensorimotor impairments of paretic upper limb correlates with activities of daily living in subjects with chronic stroke. *S. Afr. J. Physiother.* **2011**, *67*, 9–16. [CrossRef]

3. Crichton, S.L.; Bray, B.D.; McKevitt, C.; Rudd, A.G.; Wolfe, C.D. Patient outcomes up to 15 years after stroke: Survival, disability, quality of life, cognition and mental health. *J. Neurol. Neurosurg. Psychiatry* **2016**, *87*, 1091–1098. [CrossRef] [PubMed]

4. Mayo, N.E.; Wood-Dauphinee, S.; Côté, R.; Durcan, L.; Carlton, J. Activity, participation, and quality of life 6 months poststroke. *Arch. Phys. Med. Rehabil.* **2002**, *83*, 1035–1042. [CrossRef] [PubMed]

5. Mast, B.T.; Azar, A.R.; MacNeill, S.E.; Lichtenberg, P.A. Depression and Activities of Daily Living Predict Rehospitalization within 6 Months of Discharge From Geriatric Rehabilitation. *Rehabil. Psychol.* **2004**, *49*, 219. [CrossRef]

6. Sager, M.A.; Franke, T.; Inouye, S.K.; Landefeld, C.S.; Morgan, T.M.; Rudberg, M.A.; Sebens, H.; Winograd, C.H. Functional outcomes of acute medical illness and hospitalization in older persons. *Arch. Intern. Med.* **1996**, *156*, 645–652. [CrossRef]

7. Cain, A.E.; Depp, C.A.; Jeste, D.V. Ecological momentary assessment in aging research: A critical review. *J. Psychiatr. Res.* **2009**, *43*, 987–996. [CrossRef] [PubMed]

8. Troiano, R.P.; Berrigan, D.; Dodd, K.W.; Mâsse, L.C.; Tilert, T.; McDowell, M. Physical activity in the United States measured by accelerometer. *Med. Sci. Sports Exerc.* **2008**, *40*, 181–188. [CrossRef] [PubMed]

9. Trull, T.J.; Ebner-Priemer, U. Ambulatory assessment. *Annu. Rev. Clin. Psychol.* **2013**, *9*, 151–176. [CrossRef]

10. Magaziner, J.; Simonsick, E.M.; Kashner, T.M.; Hebel, J.R. Patient-proxy response comparability on measures of patient health and functional status. *J. Clin. Epidemiol.* **1988**, *41*, 1065–1074. [CrossRef]

11. Rubenstein, L.Z.; Schairer, C.; Wieland, G.D.; Kane, R. Systematic biases in functional status assessment of elderly adults: Effects of different data sources. *J. Gerontol.* **1984**, *39*, 686–691. [CrossRef] [PubMed]

12. Galperin, I.; Hillel, I.; Del Din, S.; Bekkers, E.M.J.; Nieuwboer, A.; Abbruzzese, G.; Avanzino, L.; Nieuwhof, F.; Bloem, B.R.; Rochester, L.; et al. Associations between daily-living physical activity and laboratory-based assessments of motor severity in patients with falls and Parkinson's disease. *Parkinsonism Relat. Disord.* **2019**, *62*, 85–90. [CrossRef] [PubMed]

13. Sim, I. Mobile Devices and Health. *N. Engl. J. Med.* **2019**, *381*, 956–968. [CrossRef] [PubMed]

14. Aggarwal, J.K.; Xia, L. Human activity recognition from 3d data: A review. *Pattern Recognit. Lett.* **2014**, *48*, 70–80. [CrossRef]

15. Babiker, M.; Khalifa, O.O.; Htike, K.K.; Hassan, A.; Zaharadeen, M. Automated daily human activity recognition for video surveillance using neural network. In Proceedings of the 2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), Putrajaya, Malaysia, 28–30 November 2017; pp. 1–5.

16. Zhang, S.; Wei, Z.; Nie, J.; Huang, L.; Wang, S.; Li, Z. A Review on Human Activity Recognition Using Vision-Based Method. *J. Healthc. Eng.* **2017**, *2017*, 3090343. [CrossRef]

17. Ranasinghe, S.; Al Machot, F.; Mayr, H.C. A review on applications of activity recognition systems with regard to performance and evaluation. *Int. J. Distrib. Sens. Netw.* **2016**, *12*, 1550147716665520. [CrossRef]

18. Bao, L.; Intille, S.S. Activity recognition from user-annotated acceleration data. In *International Conference on Pervasive Computing*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 1–17.

19. Nweke, H.F.; Teh, Y.W.; Mujtaba, G.; Al-Garadi, M.A. Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Inf. Fusion* **2019**, *46*, 147–170. [CrossRef]

20. Twomey, N.; Diethe, T.; Fafoutis, X.; Elsts, A.; McConville, R.; Flach, P.; Craddock, I. A comprehensive study of activity recognition using accelerometers. *Informatics* **2018**, *5*, 27. [CrossRef]

21. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognit. Lett.* **2019**, *119*, 3–11. [CrossRef]

22. Lehman, L.H.; Mark, R.G.; Nemati, S. A Model-Based Machine Learning Approach to Probing Autonomic Regulation from Nonstationary Vital-Sign Time Series. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 56–66. [CrossRef]

23. Alghamdi, M.; Al-Mallah, M.; Keteyian, S.; Brawner, C.; Ehrman, J.; Sakr, S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLoS ONE* **2017**, *12*, e0179805. [CrossRef]

24. Tan, X.; Su, S.; Huang, Z.; Guo, X.; Zuo, Z.; Sun, X.; Li, L. Wireless Sensor Networks Intrusion Detection Based on SMOTE and the Random Forest Algorithm. *Sensors* **2019**, *19*, 203. [CrossRef] [PubMed]

25. Wang, Q.; Luo, Z.; Huang, J.; Feng, Y.; Liu, Z. A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM. *Comput. Intell. Neurosci.* **2017**, *2017*, 1827016. [CrossRef] [PubMed]

26. Lee, S.-I.; Lundberg, S.M. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Long Beach, CA, USA, 2017; pp. 4765–4774.

27. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]

28. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

29. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: Stanford, CA, USA, 2009.

30. Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; Hutter, F. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Montreal, QC, Canada, 2015; pp. 2962–2970.

Int. J. Environ. Res. Public Health **2021**, *18*, 1634

16 of 16

31. Bergstra, J.; Yamins, D.; Cox, D.D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; Volume 28, pp. 115–123.

32. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. *Learning from Imbalanced Data Sets*; Springer: Cham, Switzerland, 2018.

33. Cheung, V.H.; Gray, L.; Karunanithi, M. Review of accelerometry for determining daily activity among elderly patients. *Arch. Phys. Med. Rehabil.* **2011**, *92*, 998–1014. [CrossRef] [PubMed]

34. Capela, N.A.; Lemaire, E.D.; Baddour, N. Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients. *PLoS ONE* **2015**, *10*, e0124414. [CrossRef]

35. Capela, N.A.; Lemaire, E.D.; Baddour, N.; Rudolf, M.; Goljar, N.; Burger, H. Evaluation of a smartphone human activity recognition application with able-bodied and stroke participants. *J. Neuroeng. Rehabil.* **2016**, *13*, 5. [CrossRef]

36. Bailey, R.R.; Birkenmeier, R.L.; Lang, C.E. Real-world affected upper limb activity in chronic stroke: An examination of potential modifying factors. *Top. Stroke Rehabil.* **2015**, *22*, 26–33. [CrossRef]

37. Lang, C.E.; Bland, M.D.; Bailey, R.R.; Schaefer, S.Y.; Birkenmeier, R.L. Assessment of upper extremity impairment, function, and activity after stroke: Foundations for clinical decision making. *J. Hand Ther.* **2013**, *26*, 104–114; quiz 115. [CrossRef]

38. Bailey, R.R.; Klaesner, J.W.; Lang, C.E. An accelerometry-based methodology for assessment of real-world bilateral upper extremity activity. *PLoS ONE* **2014**, *9*, e103135. [CrossRef]