Use of Accelerometry for Long Term Monitoring of Stroke Patients

ALFREDO LUCAS [ID]1, JOHN HERMIZ [ID]2, JAMIE LABUZETTA3, YEVGENIY ARABADZHI2, NAVAZ KARANJIA 3, AND VIKASH GILJA 2

[1]Department of Bioengineering, University of California, San Diego, CA 92106, USA
[2]Department of Electrical and Computer Engineering, University of California, San Diego, CA 92106, USA
[3]Department of Neurosciences, University of California, San Diego, CA 92106, USA

(Alfredo Lucas and John Hermiz contributed equally to this work.) CORRESPONDING AUTHOR: V. GILJA (vgilja@eng.ucsd.edu)

ABSTRACT Stroke patients are monitored hourly by physicians and nurses in an attempt to better understand their physical state. To quantify the patients' level of mobility, hourly movement (i.e. motor) assessment scores are performed, which can be taxing and time-consuming for nurses and physicians. In this paper, we attempt to find a correlation between patient motor scores and continuous accelerometer data recorded in subjects who are unilaterally impaired due to stroke. The accelerometers were placed on both upper and lower extremities of four severely unilaterally impaired patients and their movements were recorded continuously for 7 to 14 days. Features that incorporate movement smoothness, strength, and characteristic movement patterns were extracted from the accelerometers using time-frequency analysis. Support vector classifiers were trained with the extracted features to test the ability of the long term accelerometer recordings in predicting dependent and antigravity sides, and significantly above baseline performance was obtained in most instances ($P < 0.05$). Finally, a leave-one-subject-out approach was carried out to assess the generalizability of the proposed methodology, and above baseline performance was obtained in two out of the three tested subjects. The methodology presented in this paper provides a simple, yet effective approach to perform long term motor assessment in neurocritical care patients.

INDEX TERMS Accelerometers or wearable sensors, machine learning algorithms, neurology.

I. INTRODUCTION

Motor impairment monitoring in stroke patients admitted to the Intensive Care Unit (ICU) is crucial for understanding patient prognosis and recovery, as well as for identifying critical times for the application of medications, such as Tissue Plasminogen Activator (tPA), which can significantly improve patient outcomes [1]. It is also relevant for detection of early onset of Intensive Care Unit Acquired Weakness (ICU-AW), which can persist for up to 2 years after patient discharge [2].

The measurement of motor function is usually done through standardized measurement tests. Of particular interest for this study is the Oxford Grading system for hourly neuroassessments [3], performed by a nurse or clinical provider on the patient, whose strength is assessed and scored 0 to 5 on each limb. While such scoring is of clinical utility, it is non-ideal for multiple reasons; firstly, the exams are labor intensive and usually performed not more frequently than

every hour, leaving the possibility for major changes in motor ability to remain undetected for long periods of time. Furthermore, interobserver variability still remains a problem, where the accuracy of the assigned score is highly dependent on external factors such as the provider's expertise and time of stroke onset [4]. Finally, frequent neurological exams during a patient's hospital stay disturbs sleep and may increase delirium, which worsens morbidity and mortality [5], [6].

The use of accelerometers to monitor physical activity in critically ill subjects has been explored previously with successful results [7]. In the Neurological Intensive Care Units (Neurological ICU), the use of accelerometers for patient monitoring has been explored to detect agitation and sedation patterns [8], and to study sedentary behavior [9]. The use of accelerometers, specifically to detect changes in patient motor score in the Neurological ICU, has been explored previously utilizing the NIHSS motor score as a metric [10]–[12], as well as using different motor

TABLE 1. Subject demographics.

Subject Number	Age	Gender	Impaired Side	Length of Data Collection (days)	Minimum Limb Scores (RUE-LUE-RLE-LLE)	Maximum Limb Scores (RUE-LUE-RLE-LLE)
Subject 1	46	M	Right	7	0-1-0-1	2-5-1-5
Subject 2	39	F	Left	7	0-0-0-0	4-2-5-2
Subject 3	48	M	Left	14	1-0-1-0	4-3-4-2
Subject 4	74	M	Left	7	5-3-5-4	5-5-5-5

score scales [13]. However, these studies only performed the recordings over small time windows, with maximum recording times consisting of 10 minute epochs [10] and 12 second epochs [14] over 24 hours.

Machine learning algorithms have proven successful in characterizing motor activities extracted from accelerometers [15], [16]. Specifically, support vector machines (SVMs) have been used for movement characterization and activity recognition in accelerometers and other activity monitoring devices with surprising success rates [17]–[21]. While studies applying machine learning and big data approaches for assessment in the ICU have been explored before, they have been focused towards determining agitation and sedation patterns, and delirium state, but not motor impairment [22]–[24]. Additionally, in the studies that do apply machine learning, they do so in controlled settings, where accelerometer recordings are only done in the first days after admission and in certain controlled time windows [10].

In the present study we assess whether long term monitoring of seven days or more, using accelerometers in unilaterally impaired stroke patients in the ICU, is useful in determining motor impairment. An SVM classifier was created and trained using accelerometer derived features to classify dependent and antigravity limbs. The usefulness of the extracted features is also assessed through a recursive feature selection approach. The methods and results of this study serve as a proof of concept for the use of accelerometers as a monitoring mechanism in a challenging clinical environment, and as a way of translating accelerometry based machine learning models, widely used in other settings, into the Neurological ICU.

II. METHODS
A. CLINICAL DATA ACQUISITION
1) SUBJECT RECRUITMENT
This study was designed to be HIPAA compliant and all study procedures were reviewed and approved by the Institutional Review Board of the University of California, San Diego. Study subjects were recruited from the Neurological Intensive Care Unit at UC San Diego Medical Center - Hillcrest hospital. A total of four unilaterally impaired adult subjects were recruited, subject demographics are presented in Table 1. All subjects had experienced a stroke which led to severe unilateral impairment as determined by the practicing clinician. The subjects were enrolled in the study under their consent and remained enrolled in the study for up to 14 days

or until discharged from the Neurological ICU, whichever occurred first.

TABLE 2. Oxford motor grading scale.

Motor Score	Description
Score 0	No muscle movement
Score 1	Muscle movement without joint motion
Score 2	Moves with gravity eliminated
Score 3	Moves against gravity but not resistance
Score 4	Moves against gravity and light resistance
Score 5	Normal strength

2) MOTOR ASSESSMENT
As part of the Physical Function ICU Test (PFIT) [25], muscle movement grading was routinely performed using the Oxford Grading Motor Scale (Table 2) by the medical practitioner, and recorded on an hourly basis on the subjects' medical chart. The clinical motor score data were downloaded from the electronic medical record after removing all identifying information. Subjects 1 through 3 were recorded for the duration of their stay in the hospital without interruptions in the data collection procedure with hourly assessments. Subject 4 had some interruptions in the data collection procedure and motor scores were often obtained every 1 to 4 hours to mitigate the risk of experiencing delirium as determined by the clinical team.

Information from the Oxford Grading Motor Scale was used for this study since it has been shown to be the best estimator of antigravity muscle strength [3]. Furthermore, the Oxford Grading Scale is considered the gold standard for determining ICU acquired weakness (ICU-AW) [26], which early onset detection can aid in successful recovery post-ICU.

3) LIMB IMPAIRMENT CHARACTERIZATION
During each limb assessment a limb was classified as "dependent" if it had a motor score of 0-2 and "antigravity" if it had a motor score of 3-5. A score of 3 is chosen as a threshold since it separates those movements that can be done against gravity (scores 3, 4, and 5) and those that cannot (scores 0, 1, and 2), effectively giving a binary representation of impairment. It also allows each limb classification to contain 3 motor scores. The creation of the two groups of limbs allows the motor impairment assessment from the accelerometer features to be treated as a binary classification problem which can be analyzed with the proposed methodology. Motor score

values and impairment information for each subject is shown in Table 1.

4) ACCELEROMETER DATA COLLECTION

Upon enrollment, four tri-axial accelerometers (Axtivity AX3 Accelerometers) were placed in the left upper extremity (LUE = left arm), the right upper extremity (RUE = right arm), the left lower extremity (LLE = left leg) and the right lower extremity (RLE = right leg). All accelerometers were attached to hospital bands and placed on the subjects' wrists and ankles as shown in Figure 1. Data were continuously acquired at 100Hz for up to 14 days or until the subject was discharged from the Neurological ICU, whichever occurred first. After the study was completed, the accelerometers were removed from the extremities and the data were downloaded to a computer.
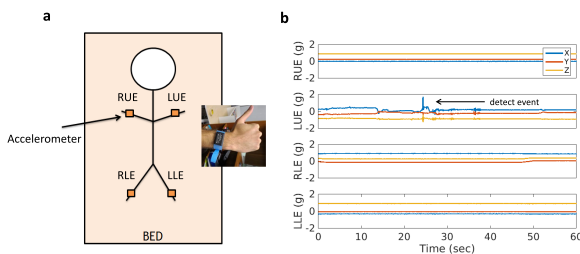

FIGURE 1. a) Illustration of experimental setup, where 4 accelerometers are mounted onto hospital bands and placed around each of the patients extremity. **b)** Raw tri-axial accelerometer measurements are performed continuously across all 4 extremities. Changes in accelerometry are detected, interpreted as movements, and counted per hour.

B. DATA PROCESSING AND ANALYSIS

All of the subsequent analyses were performed on a personal laptop computer with 16GB of RAM and an Intel Core i7 CPU with 2.80GHz running Windows 10.

1) SIGNAL PREPROCESSING AND MOVEMENT EVENT EXTRACTION

The raw accelerometer data were down-sampled from 100Hz to 50Hz, because significant signal power was only seen below 25Hz. After down-sampling, the magnitude of each accelerometer was calculated according to equation 1

$$A[t] = \sqrt{x[t]^2 + y[t]^2 + z[t]^2} \qquad (1)$$

where $A[t]$ corresponds to the magnitude of the accelerometer signal at time t and $x[t]$, $y[t]$ and $z[t]$ to the acceleration in the x, y and z direction, at time t in the accelerometers' frame of reference. Subsequently, to eliminate the baseline normal force ($9.8 \ m/s^2$), measured in the accelerometer as a constant offset, the difference between subsequent timepoints, $A'[t]$, was taken. Matlab software was used to compute features.

To classify movement events, an empirically chosen threshold was applied to the $A'[t]$ of the signal. The parameters used to detect features were manually chosen and were determined to be sufficient based on their ability to detect

events from an example window of data. Instances that lasted longer than 1 sec and had a subsequent silent period of 0.5 sec after the instance were then classified as movement events, and stored with the corresponding timestamp. All extracted movement events consisted of non-overlaping windows and were specific to each limb.

It is important to note that some of the detected movement events are not generated by the subject, but rather by clinicians who need to interact with the patient. The approach to filter these movements out is discussed in the following sections.

2) FEATURE GENERATION

The start and end timestamps for each detected event following the procedure from the previous section, were used to extract the magnitude corresponding to that time window. The time magnitude vector for each event window was subsequently used to create 9 scalar valued features. The features were selected on their ability to characterize movement on the basis of smoothness, intensity and pattern behavior, and all have been successfully used to characterize motor activity in wrist and ankle worn accelerometers [17], [27]. Furthermore, feature computation time was also considered in the selection process to ensure translational applicability.

The features extracted for each event are shown in table 3. From the time domain, the first feature consisted of the time average of the magnitude vector as shown in Equation 2,

$$A_{avg} = \frac{1}{t_f - t_i} \int_{t_i}^{t_f} A[t]dt \qquad (2)$$

for t_i and t_f corresponding to the initial and final timepoints of the event. The maximum and minimum values of the magnitude vector were also used. To characterize movement smoothness, the jerk of the signal has been previously used to characterize tremors in bradykinesia and Parkinson's disease [28], [29]. To convert the jerk into a scalar quantity the Normalized Average Rectified Jerk (NARJ) was used instead of the time average jerk since the NARJ has been shown to be a consistent metric for movement smoothing independent of signal duration [30]. The NARJ was computed as follows

$$NARJ = \frac{1}{t_f - t_i} \int_{t_i}^{t_f} \left| \frac{dA[t]}{dt} \right| \qquad (3)$$

An FFT transform of the magnitude vector was used to obtain frequency based features. The first and second
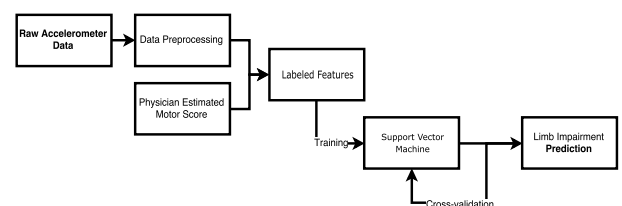

FIGURE 2. Flowchart presenting the pipeline used to generate the results.

dominant frequencies (excluding DC) and their corresponding FFT power were extracted as features, all of which have been proven successful for accelerometer based movement characterization [27]. The use of multiple combinations of these features have also proven useful, but with minimal improvements in the classification accuracy, therefore they were excluded as a means of minimizing the number of features [17].

After the above scalar features were computed for every extracted event, the average of each feature for events taking place in a single hour was computed, resulting in one feature vector for every limb and for every hour. This allowed us to include a final feature that consisted of the number of events that took place in that given hour. The reason for combining event features in an hourly fashion is twofold. First, since motor assessments are performed every hour, the clinical score between assessments is not accurately known and assuming that every event within the hour has the same score as the hour itself would be an oversimplification. Secondly, some of the registered events will correspond to movements induced by interactions between the practitioner and the subject. Under the assumption that those interactions only happen a small number of times within the hour, taking the hourly average of the features will allow to minimize the effects these confounding movements have on the performance of the system.

The features were organized into a matrix where each row corresponds to an hour where events were recorded and columns correspond to the respective features for that hour. A given row of the feature matrix was labeled to be from a dependent limb if the motor score for that limb at that hour was less than 3, and antigravity otherwise, in keeping with the previously defined thresholds. The columns of the feature matrix (feature vectors) were scaled to have zero mean and unit variance, which is necessary to prevent uneven feature scaling to affect the results. This normalization was done per subject, per day. That is, the feature vectors corresponding to the same subject for a given day were normalized under the same distribution to zero mean and unit variance.

3) SUPPORT VECTOR MACHINE CLASSIFIER

The use of Support Vector Machines for the characterization of movements using accelerometers has been successfully explored before [17]–[21]. Furthermore, SVM's suitability for binary classification problems with small numbers of features makes it an ideal choice of algorithm for this study [31].

The open source Python library *scikit-learn* [32] was used, together with Python 3.6 to apply the classification analysis to the data. Support Vector Classifiers (SVCs) were trained to classify between dependent and antigravity limbs. In its essence SVCs project the input feature space into a higher dimensional features space in which a hyperplane described by "support vectors" is used to classify the data [33]. A linear kernel was chosen for all the trained SVCs. The choice of a simple linear kernel prevents overfitting of the training set in the presence of small datasets. The two hyperparameters

TABLE 3. Accelerometer derived features.

Name	Description
Magnitude Average	- Time average of the magnitude vector
Maximum Magnitude	- Maximum value of the magnitude vector
Minimum Magnitude	- Minimum value of the magnitude vector
NARJ	- Normalized average rectified jerk [30]
Power 1	- Power of the first dominant FFT coefficient
Power 2	- Power of the first dominant FFT coefficient
Frequency 1	- Frequency corresponding to the first dominant FFT coefficient
Frequency 2	- Frequency corresponding to the second dominant FFT coefficient
Movement Count	- Number of events recorded in the hour

of the classifier, C and γ were determined using a coarse parameter grid search with values $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ and $C = 2^{-15}, 2^{-13}, \ldots, 2^{3}$, according to the procedure described in [34]. The optimal parameters obtained from the grid search were $C = 2$ and $\gamma = 0.5$.

SVCs were trained for each subject and separately for upper and lower extremities within each subject. The data were divided using an 80/20 approach, where 80% of the data is used for training and validation and 20% is used for testing. Testing and training sets were ensured to have similar ratios of dependent and antigravity instances. This process was repeated for every patient and type of extremity (upper and lower). In order to evaluate the ability of cross-limb information in helping to classify dependent limbs, a third combined classifier was created for each subject, in which training and prediction occurred in combined upper and lower limbs. The feature selection process and cross-validation was identical.

To determine the validity of a classification, the probability output from the classifier was used. In general, if the posterior probability of an instance belonging to the dependent class is over 50%, the SVC will assign that class to the tested instance. However, since the classifier is not perfect, there will be a certain degree of uncertainty with each prediction, which can be estimated by the closeness of a given prediction to the decision boundary. As a means of increasing the certainty of the predictions, those that had probabilities within a region close to the decision boundary were labeled as uncertain. To determine this region, the probability of an instance belonging to the dependent class was used. Since the clinical risk of a false negative (incorrectly predicting a dependent limb as antigravity) is larger than the risk of a false positive (incorrectly predicting an antigravity limb as dependent), the lower bound of the probability region was set to 42% while the upper bound was set to 54%. That is, for

an instance to be classified as antigravity, the probability of it belonging to the dependent class must be smaller than 42%, while for it to be classified as dependent, the probability of it belonging to the dependent class must be larger than 54%. Instances with probabilities between 42% and 54% are otherwise labeled as uncertain. These bounds were chosen such that the number of instances labeled as uncertain was less than 20% of the tested instances.

Maintaining the dependent and antigravity labels of the feature matrix as a template, the entries of the matrix were populated with normalized random data between 0 and 1 to train a baseline classifier. The random feature matrix was subject to the same scaling as the actual feature matrix and used to train a SVC. Cross-validation scores were computed for the baseline classifier and used as the baseline accuracy for its corresponding actual classifier. Given that the number of cross validation accuracy scores is only 10 for every actual classifier, the normality of this data cannot be accurately determined. For this reason a nonparametric Wilcoxon rank-sum test was used to compare the cross-validation scores of the actual classifier with the accuracy of the baseline classifier. The statistical significance of the accuracy values for each trained classifier was calculated by subtracting the cross-validation accuracy scores from their corresponding baseline accuracy and performing a one sided Wilcoxon rank-sum test in the resulting datapoints.

Only subjects 1 through 3 were used for creating the individual classifiers since subject 4's minimum motor score on the dependent side was 3, therefore all of its features were labeled as antigravity.

4) FEATURE RELEVANCE ASSESSMENT

In order to validate the choice of features, a recursive feature selection approach was taken. Given the complete set of features described previously, an SVC was trained using all possible combinations of features in sets ranging from 1 feature to 9 features. Each individual classifier previously described (upper, lower and combined extremity classifiers), for subjects 1 through 3 was used. With the same specifications described in the previous section, a 10-fold cross-validation was applied to each classifier and the average cross-validation accuracy for all folds and all classifiers was used as a metric for the performance of each feature set. The test set did not enter the feature selection process in any way.

5) LEAVE-ONE-SUBJECT-OUT APPROACH

In order to assess the generalizability of the proposed methodology, a leave-one-subject-out approach was used. For this approach, all the data from a single subject were excluded from the training set and then the model was tested on that excluded data. The SVCs used in this approach had the same hyperparameters and linear kernel type as in the previously described classifiers. To assess which features were generalizable across subjects, a leave-one-subject-out approach leaving subject 3 out, and training on subjects 1, 2, and 4, was performed. Using a similar approach as that shown in

the previous section, the performance in the left out test set after training with different feature sets was assessed for a classifier using only information from the upper extremities, the lower extremities, and combined upper and lower extremities. After the features with the best performance were identified in subject 3, a final leave-one-subject-out approach training on subjects 2, 3 and 4, and testing on subject 1 was carried out. For comparison, a baseline classifier was also trained following the same approach described previously, and tested on the data of subject 1. Subject 4 was not assessed for the leave-one-subject-out approach due to the absence of any impaired limbs according to the specified criteria. Subject 2 was assessed for the leave-one-subject-out approach by training on subjects 1, 3, and 4, however, it seemed that none of the features generalized well as performance was no better than baseline.

III. RESULTS

The results from the recursive feature selection are shown in Figure 3. It can be seen that every feature used yields above baseline mean cross validation accuracy. The largest accuracy is achieved with classifiers trained using all available features as shown by the blue triangle, and with a classifier using all features except the *average magnitude*, which is represented by highest red circle. A general downwards trend is observed as less features are used for training the model, and there seems to be a clear separation between certain sets of features, as represented by a large vertical gap in the figure. The single features with the largest accuracy were the *average magnitude*, *movement count* and *power 2*. Furthermore, feature sets ranging between 2 and 7 features containing either of these three features always appeared above the vertical gap suggesting the relative importance of these three features. The individual classifiers were trained using all features as the difference in cross-validation performance between the highest performing 8 feature set and the complete feature set was negligible.

FIGURE 3. Scatterplot showing the recursive feature sets mean cross-validation accuracy across all individual classifiers. Each data point is color coded to the number of features in the feature set it represents. The number of features in each set decreases as the points go to the right. Mean baseline classifier is shown with a dashed line.
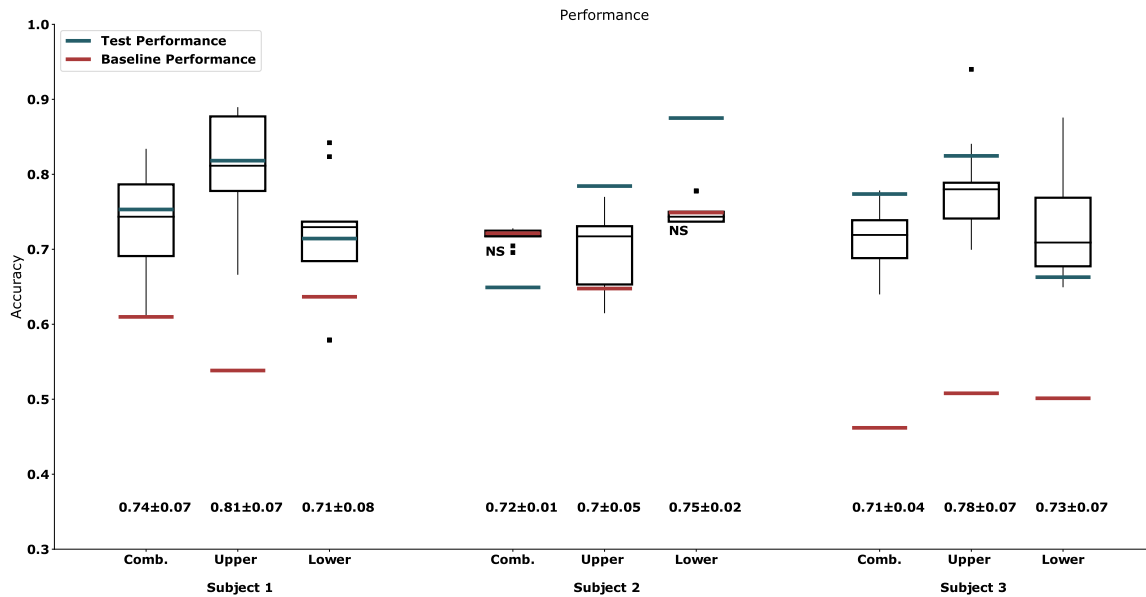
FIGURE 4. Boxplot representing the accuracy obtained after a 10-fold cross-validation for each limb classifier and the combined classifier for each subject. Baseline obtained values are shown as a red line and test set accuracy is shown as a blue line. Outliers are shown as black squares. Mean± standard deviation for all boxplots shown underneath. Not statistically significant differences between the cross-validation results and the baseline classifier are denoted by NS.

The results from the trained SVCs are shown in Figure 4. The combined classifiers, across all trained subjects, had on average a cross-validation accuracy of 0.72 ± 0.05 and a test set accuracy of 0.73 ± 0.05, the upper classifiers an average cross-validation accuracy of 0.76 ± 0.08 and a test set accuracy of 0.81 ± 0.02, and the lower classifiers an average a cross-validation accuracy of 0.78 ± 0.13 and a test set accuracy of 0.80 ± 0.06. Most instances had cross-validation performance statistically significant ($P < 0.05$) above that of the baseline classifier, with the exception of the combined ($P = 0.25$) and lower ($P = 0.39$) extremity classifiers for subject 2. Furthermore, all test set accuracies were within the cross-validation scores range and above baseline, with the exception of the combined classifier in subject 2.

The feature selection process showed that the features that performed the best in the leave-one-subject-out approach were the *maximum magnitude*, *power 2* and the *movement count*. The leave-one-subject-out approach on subject 3, which was used for feature selection, yielded test set accuracies of 0.66, 0.77 and 0.73, with baseline accuracies of 0.51, 0.51 and 0.49, for combined, upper and lower extremity classifiers respectively. The leave-one-subject-out approach on subject 1, had accuracies of 0.74, 0.82 and 0.72, and baseline accuracies of 0.61, 0.56 and 0.65 for combined, upper and lower extremity classifiers respectively. Finally the leave-one-subject-out approach on subject 2, had accuracies of 0.50, 0.51 and 0.47, and baseline accuracies of 0.35, 0.61 and 0.74 for combined, upper and lower extremity classifiers respectively. These results are summarized in Table 4 The confusion matrices and ROC curves for the leave-one-subject-out approach on subject 1 are shown in Figure 5. The confusion matrices shown in Figure 5A-C have the

TABLE 4. Leave-one-subject-out accuracy for subjects 1 through 3, and all corresponding classifiers. Baseline accuracy is shown in parenthesis.

	Combined	Upper	Lower
Subject 1	0.74 (0.61)	0.82 (0.56)	0.72 (0.65)
Subject 2	0.50 (0.35)	0.51 (0.61)	0.47 (0.74)
Subject 3	0.66 (0.51)	0.77 (0.51)	0.73 (0.49)

largest values in their diagonals, with a large number of true negatives (antigravity instances classified as antigravity), followed by a smaller number of true positives (dependent instances classified as dependent). In the off-diagonals, there was a larger number of false negatives (dependent instances classified as antigravity), than false positives (antigravity limbs classified as dependent), and this is consistent across all matrices. The receiver operator characteristics (ROC) curves for combined, upper and lower extremity classifiers are also shown in Figure 5D-F. The area under the ROC curves was largest for the upper extremity trained classifier with a value of 0.87, followed by the combined classifier with a value of 0.76 and the lower extremity classifier with a value of 0.74. For the leave-one-subject-out approach of subject 1, 67% of the incorrect classifications took place in the second half of the dataset, that is, in the latter days in which the subject was monitored. For the leave-one-subject-out approach of subject 3, the errors were distributed equally amongst the two halves of the dataset.

IV. DISCUSSION

The results presented in this study successfully demonstrate the ability of features extracted from continuous accelerometer recordings in the NeuroICU to determine gravity and
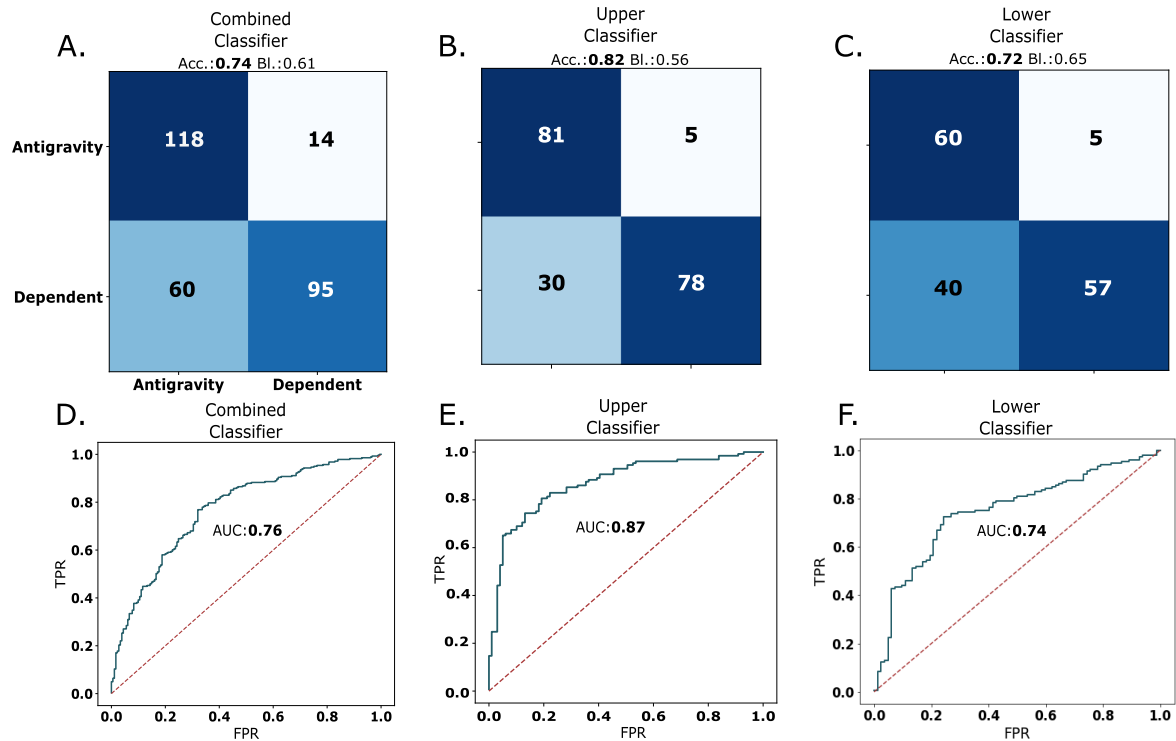
FIGURE 5. Confusion matrices for the A. combined, B. upper and C. lower extremity classifiers for the final leave-one-subject-out approach tested in subject 1. The classifier accuracy (Acc.) and baseline classifier accuracy (Bl.) are shown at the top of the confusion matrix. The corresponding ROC curves for the D. combined, E. upper and C. lower extremity classifiers are shown in blue. The dashed line represents the ideal ROC for a completely random classifier. The area under the ROC curve (AUC) is also shown .

antigravity limbs. Of primary importance is that this is the first study, to our knowledge, that has attempted to assess motor impairment in the ICU for 7 days or more in individual subjects. We have also demonstrated that the proposed methodology is capable of generalizing to new subjects with minimal modifications, allowing for a simple, yet effective, way of performing motor assessment in the NeuroICU. While the sample size is limited, the results should serve as a lower bound for the performance that future studies should aim to achieve.

The iterative feature assessment shown in Figure 3 demonstrates that all the features extracted from the accelerometer are informative. The general downward trend as the number of features is decreased suggests that while certain features might dominate in importance, such as the *average magnitude* and the *movement count*, the contribution of the other features is sufficient to yield larger accuracies when they are included in the model. Simply using the movement count, as traditionally done in approaches involving actigraphy [35], has been shown to be ineffective in an ICU setting [36]. The incorporation of these new features attempts to account for other characteristics of the subject's movement such as smoothness and idiosyncratic movement patterns, which might explain the additional accuracy obtained by including these features. Additionally, their computation is fast and straightforward. After extraction of all the movement

events with the proposed approach, the computation of the entire feature matrix for a given subject takes on average 3 seconds, resulting in about 0.5 milliseconds to compute all the features for a single event. The relative simplicity in terms of their computation makes them ideal for mobile or portable applications. Since they are purely derived from accelerometer information, they can be easily incorporated into digital actigraphs to improve their effectiveness in ICU settings [36]. Furthermore, all of the proposed methodology involved easily accessible resources such as off-the-shelf accelerometers and open source software, which significantly increases the accessibility of the proposed approach.

The results of the support vector classifiers (Figure 4) also proved to be promising. Despite certain instances where the test set performance was sub-optimal, such as in the combined limb classifier of subject 2, the results from the cross validations seem to generalize well into the testing sets. Furthermore, most cross-validation accuracy measurements were statistically above those of the baseline classifiers, with differences in accuracy of up to 0.35 in some cases. Additionally, the training of each classifier, after the feature matrix had been constructed and filtered, ranged between 20 to 100 milliseconds, once again proving to be sufficiently fast for portable applications.

In general, the results shown in Figure 4 suggest that the performance of the combined classifier is equivalent or

better than that of the lower classifiers, and equivalent or slightly inferior to the upper classifier. These cross-limb combinations have not been explored in the literature for ICU monitoring, likely due to the expected differences between movements in the upper and lower limbs, which are likely to be exacerbated in bed ridden individuals. However, since these individuals are unilaterally impaired, the degree of mobility between ipsilateral limbs might be sufficiently similar to successfully train a model using information from all limbs. These results show that models trained on information from both upper and lower limbs in unilaterally impaired patients in the Neurological ICU can be successfully used for long term hourly limb classification.

The leave-one-subject-out approach results are also promising. With a maximum test set accuracy of 0.82 in the upper limb classifier, and an average accuracy of 0.76 across all classifiers, it can be seen that the performance of this approach is similar to the individually trained classifiers. Furthermore, the features that performed well in this approach encompass the strength, the smoothness and the frequency (as in movement counts) of the movements in each subject. The *maximum magnitude* directly relates to the acceleration of the movement, as it determines the upper limit of that acceleration, and therefore it is expected to directly relate to the strength of the subject. The *power 2* relates to the smoothness of the movements, as it describes the relevance of the second dominant frequency. These two features, combined with the *movement count*, roughly provide the same information as all the other features used for the individual classifiers. As shown in the recursive feature selection for the individual classifiers the movement count, power 2 and the average magnitude contributed positively to the performance of the classifiers. Therefore, it is not surprising that two of these three features, the movement count and the power 2, are able to generalize well across subjects. The third feature, the maximum magnitude, is closely related to the average magnitude, therefore it is also consistent with the previous discussion. The fact that the entire feature set does not perform as well as the reduced feature set in the leave-one-subject-out approach might be evidence that the entire feature set might be causing the model to over-fit to individual subjects. This can be useful in cases when individualized models are desired, such as in the individual classifiers that were studied. In the case of a generalized model, however, a small subset of features capable of encompassing the diverse movement dynamics of multiple subjects, such as the ones chosen here, seems to prove more useful.

The confusion matrices shown in Figure 5A-C also show that in general, the model is highly capable of determining when a limb belongs to the antigravity class, while it struggles more in distinguishing dependent limbs. This behavior is consistent across classifiers, and might be due to the decreased number of dependent training examples due to the nature of the hemiparesis. Nevertheless, the results still suggest that the model has the capacity to classify dependent limbs with a performance above random guessing. The distribution of

the errors seems to be slightly skewed towards the latter part of the dataset in subject 1, and equally distributed in subject 3. This is promising, as it suggests that the model performs very similar, at least in subject 3 and to a lesser extent in subject 1, at the beginning and end of the subject's stay in the ICU. This can serve as potential evidence of the ability of the proposed methodology to generalize well in long term Neurological ICU monitoring. While still below the 90% multi-class accuracy presented in [10], these results are tested in more than 130 instances, while previous studies only do so in 5 or less. While the generalization of the model only worked in two the three subjects tested in the leave-one-subject-out approach, the results are still promising since both subjects were very different in terms of the study. Subject 1 was right side impaired and subject 3 was left side impaired, and both had very different distribution of motor scores and lengths of stay. Despite not testing subject 4 in the leave one out approach, training with data from subject 4 seemed to have contributed to the performance, as ignoring it from the training set caused decreased accuracy when testing on subjects 1 and 3. Similarly, despite the poor performance when testing on subject 2, training with data from subject 2 in the leave-one-subject-out approach of subjects 1 and 3 proved useful in terms of performance. This is indicative that the model is indeed learning from these subjects, and might suggest that there is potential for improvement if more data is provided.

The results of the SVCs are also important in the context of long term monitoring of individuals admitted to the ICU. Similar studies [7], [10], [11], [37] have only performed accelerometry recordings in small time windows and up to 3 days of interrupted monitoring, and to our knowledge, this is the only study that has obtained this performance using long term recordings. Since every SVC is trained and tested using information from the entirety of the subject's stay in the ICU, the significantly above baseline performance of these classifiers suggests that they are flexible enough to make accurate predictions in situations where a limb might transition from dependent to antigravity, as it was the case in initially dependent limbs of subjects 1, 2 and 3 (Table 1).

A clear limitation of this study is the sample size, and despite our results using a leave-one-subject-out approach might suggest a potential for generalizability, a larger sample size is needed for a definite conclusion. However, given the absence of a study with longer than 24 hour continuous accelerometry based monitoring in ICU settings, the results from this study serve as a lower bound for the performance, and as a proof of concept of a potential course of action. Another important limitation is the presence of confounding movements induced by clinical practitioner-patient interactions. While efforts were made to mitigate the effects of these movements, more effective filtering approaches such as video monitoring could prove more useful and improve upon the results. An alternative strategy is to use video alone to monitor activity. Previous work has demonstrated the ability to track upper body joints in the Epilepsy Monitoring Unit from RGB-

video [38]. This approach is advantageous because sensors do not touch the body, which removes the risk of skin irritation and other compilations such as neglecting to remove non-MRI compatible sensors prior to MRI. There are many approaches and opportunities to capture patient movement information that may useful for determining neurological state. Future studies should also aim at performing a finer motor assessment that is not only hourly, but includes assessments in smaller time windows in order to achieve a finer temporal resolution. This can aid in detecting important changes earlier, such that the proper intervention can take place. A final limitation is the susceptibility of the ground truth labels, namely the clinician assigned motor scores, to individual bias. Despite efforts to standardize the way hourly clinical assessments are performed, inter and intra-observer variability still serve as confounding factors that limit the accuracy of a given score, especially in overcrowded hospitals, where the time per patient needs to be minimized and mistakes are more likely. This study should serve as one of many starting points to develop consistent and objective ways to monitor motor impairment in the Neurological ICU.

V. CONCLUSION

The present work served to explore the use accelerometry for long term monitoring of severe motor impairment in unilaterally impaired subjects in the Neurological ICU.We have shown that the movement information obtained from the accelerometers can be used to create informative features that can potentially be used in new monitoring approaches. We have also shown that Support Vector Classifiers are capable of classifying dependent and antigravity limbs with above baseline performance using solely movement information extracted from the accelerometers. The incorporation of a leave-one-subject-out approach shows that accelerometer information acquired from different subjects can be useful in training classifiers that generalize to new subjects for long term ICU monitoring. The proposed approaches serve as an initial proof of concept for the use of accelerometers as a long term monitoring mechanism in a challenging clinical environment such as the Neurological Intensive Care Unit.

REFERENCES

[1] W. M. Clark *et al.*, "Recombinant tissue-type plasminogen activator (Alteplase) for ischemic stroke 3 to 5 hours after symptom onset: The ATLANTIS study: A randomized controlled trial," *JAMA*, vol. 282, no. 21, pp. 2019–2026, Dec. 1999.

[2] G. Hermans and G. van den Berghe, "Clinical review: Intensive care unit acquired weakness," *Critical Care*, vol. 19, no. 1, p. 274, Dec. 2015.

[3] H. J. Hislop, D. Avers, M. Brown, and L. Daniels, *Daniels Worthingham's Muscle Testing: Techniques of Manual Examination and Performance Testing*. St. Louis, MO, USA: Elsevier, 2014.

[4] P. J. Hand *et al.*, "Interobserver agreement for the bedside clinical assessment of suspected stroke," *Stroke*, vol. 37, no. 3, pp. 776–780, Mar. 2006.

[5] R. E. Salas and C. E. Gamaldo, "Adverse effects of sleep deprivation in the ICU," *Crit. Care Clinics*, vol. 24, no. 3, pp. 461–476, Jul. 2008.

[6] P. Pandharipande *et al.*, "Long-term cognitive impairment after critical illness," *New England J. Med.*, vol. 369, no. 14, pp. 1306–1316, Oct. 2013.

[7] A. C. Verceles and E. R. Hager, "Use of accelerometry to monitor physical activity in critically ill subjects: A systematic review," *Respiratory Care*, vol. 60, no. 9, pp. 1330–1336, Sep. 2015.

[8] R. Raj, K. Ussavarungsi, and K. Nugent, "Accelerometer-based devices can be used to monitor sedation/agitation in the intensive care unit," *J. Crit. Care*, vol. 29, no. 5, pp. 748–752, May 2014.

[9] A. E. Mattlage, S. A. Redlin, M. A. Rippee, M. G. Abraham, M. M. Rymer, and S. A. Billinger, "Use of accelerometers to examine sedentary time on an acute stroke unit," *J. Neurologic Phys. Therapy*, vol. 39, no. 3, pp. 166–171, Jul. 2015.

[10] D. Kumar, J. Gubbi, B. Yan, and M. Palaniswami, "Motor recovery monitoring in post acute stroke patients using wireless accelerometer and cross-correlation," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jan. 2013, pp. 6693–6703.

[11] J. Gubbi, A. S. Rao, K. Fang, B. Yan, and M. Palaniswami, "Motor recovery monitoring using acceleration measurements in post acute stroke patients," *Biomed. Eng. Online*, vol. 12, no. 1, p. 33, Dec. 2013.

[12] C. L. Heron *et al.*, "Wireless accelerometry is feasible in acute monitoring of upper limb motor recovery after ischemic stroke," *Cerebrovascular Diseases*, vol. 37, no. 5, pp. 336–341, Jan. 2014.

[13] S. C. van der Pas, J. A. Verbunt, D. E. Breukelaar, R. van Woerden, and H. A. Seelen, "Assessment of arm activity using triaxial accelerometry in patients with a stroke," *Archives Phys. Med. Rehabil.*, vol. 92, no. 9, pp. 1437–1442, Aug. 2011.

[14] M. J. Grap *et al.*, "Sedation in adults receiving mechanical ventilation: Physiological and comfort outcomes," *Amer. J. Crit. Care*, vol. 21, no. 3, pp. e53–e64, May 2012.

[15] A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, Feb. 2010.

[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. 5, pp. 31314–31338, Dec. 2015.

[17] A. Mannini, S. S. Intille, M. Rosenberger, A. M. Sabatini, and W. Haskell, "Activity recognition using a single accelerometer placed at the wrist or ankle," *Med. Sci. Sports Exerc.*, vol. 45, no. 11, pp. 2193–2203, Nov. 2013.

[18] R. Alvarez, E. Pulido, and D. A. Sierra, *Climbing/Descending Stairs Detection Using Inertial Sensors and Implementing PCA and a SVM Classifier*. New York, NY, USA: Springer, 2017.

[19] M. Schmid *et al.*, "SVM versus MAP on accelerometer data to distinguish among locomotor activities executed at different speeds," Computational and mathematical methods in medicine, Epub, 2013.

[20] L. C. Wu *et al.*, "Detection of american football head impacts using biomechanical features and support vector machine classification," *Sci. Rep.*, vol. 8, p. 855, Dec. 2017.

[21] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *Proc. Int. Workshop Ambient Assisted Living*. New York, NY, USA, Springer, 2012, pp. 216–223.

[22] A. Davoudi *et al.*. (2018). "The intelligent ICU pilot study: Using artificial intelligence technology for autonomous patient monitoring." [Online]. Available: https://arxiv.org/abs/1804.10201

[23] P. Casale, O. Pujol, and P. Radeva, "Classifying agitation in sedated ICU patients," in *MICCAT Workshop Proc.*, vol. 2, Nov. 2010.

[24] S. Lee *et al.*, "Constructing a Bio-Signal Repository from an Intensive Care Unit for Effective Big-data Analysis: Poster Abstract," in *Proc. 14th ACM Conf. Embedded Netw. Sensor Syst.*, Nov. 2016, pp. 372–373.

[25] L. Denehy *et al.*, "A physical function test for use in the intensive care unit: Validity, responsiveness, and predictive utility of the physical function ICU test (Scored)," *Phys. Therapy*, vol. 93, pp. 1636–1645, Dec. 2013.

[26] J. Fuller, J. Granton, and I. McConachie, *Handbook ICU Therapy*. Cambridge, U.K.: Cambridge Univ. Press, Dec. 2014.

[27] S. Zhang, A. V. Rowlands, P. Murray, and T. L. Hurst, "Physical activity classification using the GENEA wrist-worn accelerometer," *Med. Sci. Sports Exerc.*, vol. 44, pp. 742–748, Apr. 2012.

[28] N. Bayle *et al.*, "Movement smoothness differentiates voluntary from parkinsonian bradykinesia," *J. Addiction Res. Therapy*, vol. 7, pp. 1–8, Jan. 2016.

[29] Y. Beck, T. Herman, M. Brozgol, N. Giladi, A. Mirelman, and J. M. Hausdorff, "SPARC: A new approach to quantifying gait smoothness in patients with Parkinson s disease," *J. NeuroEng. Rehabil.*, vol. 15, no. 1, p. 49, Jun. 2018.

[30] J. A. Cozens and B. B. Bhakta, "Measuring movement irregularity in the upper motor neurone syndrome using normalised average rectified jerk," *J. Electromyography Kinesiology*, vol. 13, no. 1, pp. 73–81, Feb. 2003.

[31] F. Colas and P. Brazdil, "Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks," in *Proc. IFIP Int. Conf. Artif. Intell. Theory Pract.*, vol. 217, 2006, pp. 169–178.

[32] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Mach. Learn.*, vol. 12, pp. 2825–2830, Oct. 2011.

[33] C. Cortes and V. Vapnik, "Support-vector Networks," *Mach. Learn.*, vol. 20, pp. 273–297, Sep. 1995.

[34] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci., Nat. Taiwan Univ., Taipei, Taiwan, 2016. [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[35] V. Ibáñez, J. Silva, and O. Cauli, "A survey on sleep assessment methods," *PeerJ*, vol. 6, p. e4849, May 2018.

[36] A. W. van der Kooi *et al.*, "Sleep monitoring by actigraphy in short-stay ICU patients," *Crit. Care Nursing Quart.*, vol. 36, pp. 169–173, Jun. 2013.

[37] M. Noorkõiv, H. Rodgers, and C. I. Price, "Accelerometer measurement of upper extremity movement after stroke: A systematic review of clinical studies," *J. NeuroEng. Rehabil.*, vol. 11, no. 1, p. 144, Oct. 2014.

[38] K. Chen *et al.*, "Patient-specific pose estimation in a clinical environmen," in *Proc. Socal Mach. Learn. Symp.*, Oct. 2017, pp. 1–9.