Evaluation of Machine Learning Models for Classifying Upper Extremity Exercises Using Inertial Measurement Unit-Based Kinematic Data

Andrew Hua, Pratik Chaudhari, Nicole Johnson, Joshua Quinton, Bruce Schatz, David Buchner, and Manuel E. Hernandez

***Abstract*—The amount of home-based exercise prescribed by a physical therapist is difficult to monitor. However, the integration of wearable inertial measurement unit (IMU) devices can aid in monitoring home exercise by analyzing exercise biomechanics. The objective of this study is to evaluate machine learning models for classifying nine different upper extremity exercises, based upon kinematic data captured from an IMU-based device. Fifty participants performed one compound and eight isolation exercises with their right arm. Each exercise was performed ten times for a total of 4500 trials. Joint angles were calculated using IMUs that were placed on the hand, forearm, upper arm, and torso. Various machine learning models were developed with different algorithms and train-test splits. Random forest models with flattened kinematic data as a feature had the greatest accuracy (98.6%). Using triaxial joint range of motion as the feature set resulted in decreased accuracy (91.9%) with faster speeds. Accuracy did not decrease below 90% until training size was decreased to 5% from 50%. Accuracy decreased (88.7%) when splitting data by participant. Upper extremity exercises can be classified accurately using kinematic data from a wearable IMU device. A random forest classification model was developed that quickly and accurately classified exercises. Sampling frequency and lower training splits had a modest effect on performance. When the data were split by subject stratification, larger training sizes were required for acceptable algorithm performance. These findings set the basis for more objective and accurate measurements of home-based exercise using emerging healthcare technologies.**

***Index Terms*—Biomechanics, classification, inertial measurement units, machine learning, physical therapy.**

I. Introduction

EXERCISE regimens are prescribed by physical therapists for a variety of purposes, including treatment of chronic diseases (e.g. rheumatoid arthritis), treatment of acute joint injuries (e.g. tears of rotator cuff muscles at the shoulder), and prevention of injuries (e.g. fall injuries in older adults) [1], [2]. When safe to do so, including home-based (unsupervised) exercise in a prescribed regimen is preferable. Exercise can be prescribed on most or all days of the week, making supervision by therapists expensive and logistically difficult.

However, there are challenges to attaining the prescribed quantity and quality of home exercise. One study reported that nearly half of patients assigned self-monitored physical therapy (PT) had to be switched to supervised PT after six weeks, due to muscle atrophy, reduced range of motion (ROM), and low compliance [3]. Given that higher adherence to PT exercise is associated with greater physical function, self-perceived effect, and decreased pain [4], it is concerning that other studies have reported only 35-72% of participants had complete adherence to prescribed PT exercise [5], [6]. Exercise adherence is most commonly recorded with self-report exercise logs. However, exercise logs could potentially cue patients to exercise, and as such, the findings from studies using exercise logs may not be generalizable to wider clinical practice [5], [7]. While exercise logs are commonly used to monitor home exercise, logs have limitations due to over-reporting and memory errors by patients and due to the inherent difficulty of measuring exercise quality by self-report.

Thus, it is important to develop objective methods of monitoring home exercise. One approach is to use Red, Green, Blue, and Depth (RGB-D) cameras, such as the Microsoft Kinect. RGB-D cameras are widely used for gesture recognition and motion

Manuscript received October 2, 2019; revised February 24, 2020, April 27, 2020, and May 19, 2020; accepted May 29, 2020. Date of publication June 4, 2020; date of current version September 3, 2020. This work was supported in part by the Medical Scholars Program at the University of Illinois at Urbana-Champaign by a Patricia J. and Charles C.C. O'Morchoe Fellowship in Leadership Skills Awards and a Graduate Fellowship. *(Corresponding author: Andrew Hua.)*

Andrew Hua, David Buchner, and Manuel E. Hernandez are with the Department of Kinesiology and Community Health, University of Illinois at Urbana-Champaign, Champaign, IL USA (e-mail: ahua5@illinois.edu; dbuchner@illinois.edu; mhernand@illinois.edu).

Pratik Chaudhari is with the Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL USA (e-mail: pgc2@illinois.edu).

Nicole Johnson is with the Department of Bioengineering, University of Illinois at Urbana-Champaign, Champaign, IL USA (e-mail: nicolej2@illinois.edu).

Joshua Quinton is with the Department of Physics, University of Illinois at Urbana-Champaign, Champaign, IL USA (e-mail: jquinto2@illinois.edu).

Bruce Schatz is with the Department of Bioengineering and Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Champaign, IL USA (e-mail: schatz@illinois.edu).

Digital Object Identifier 10.1109/JBHI.2020.2999902

capture [8], [9] and are capable of high multi-classification performance with accuracies exceeding 90% [8], [10], [11]. Reflexion Health, Inc., has developed a product called Vera, which is capable of accurately counting exercise repetitions as well as differentiating between acceptable and unacceptable form [12]. However, RGB-D cameras have a few limitations, including limited portability, accuracy, and significant physical space requirements. To use an RGB-D camera in the home, adequate space between the user and camera needs to be provided (e.g. the Microsoft Kinect requires at least six to eight feet of open space), and portability of the cameras will be limited due to the size of the camera, computers, and power cables. Measures from RGB-D cameras can deviate as much as 16° from the actual patient's movement [13], raising questions about their ability to monitor exercise in patients at increased risk of injury or re-injury with excessive joint movement [14].

A second approach uses small, inexpensive sensors called IMUs (inertial measurement units), which measure orientation in space. By use of special clothing to secure the IMUs to the body, the IMUs measure body movements. IMUs have been used in athletes to monitor movements related to risk of overuse injuries, such as the number of baseballs thrown or volleyballs served [15]. Xsens Technologies B.V. (Netherlands) has developed IMU-based motion capture systems for research use [13], [14], while others are developing lower cost systems intended for clinical use [15]– [17]. Studies report that IMUs can be used to differentiate correct versus incorrect exercise form [16], [18], [19]. A small study of two-weeks of PT exercise using an IMU device reported high adherence to exercise and positive user experience, even though the device had bugs and inaccuracies [20]. More research using IMU-based systems is appropriate, as these systems may provide an inexpensive, portable, accurate, and adaptable technology for use in home settings. With automatic activity logging, combined with the internet of things (IoT), messages can be sent through smartphones or smart homes to remind patients to complete their daily exercises, which has been shown to improve exercise adherence [21]. These systems also offer clinicians the ability to remotely monitor patients and identify patients in need of more intensive rehabilitation, so as to realize more personalized and precise healthcare.

We propose a low-cost IMU-based device that can be worn by PT patients during exercise for the purpose of monitoring exercise. The device costs less than $150 and consists of a Raspberry Pi 3 Model B and 4 Adafruit BNO055 9DOF IMUs. In part, the shoulder was chosen due to the lack of IMU-related research on upper extremity movement [22]. The objectives of this study are: 1) to evaluate machine learning models for classifying nine different upper extremity exercises, based upon biomechanics captured from the IMU-based device, and 2) to determine the effect of various train-test splits and sampling frequencies (64 Hz, 11 Hz, and 5 Hz) of the IMU device on classifier performance. As high accuracy (>89%) has been reported when using a smartwatch to classify shoulder exercises [23], we sought to exceed a 90% classification accuracy as our device incorporated multiple sensors.

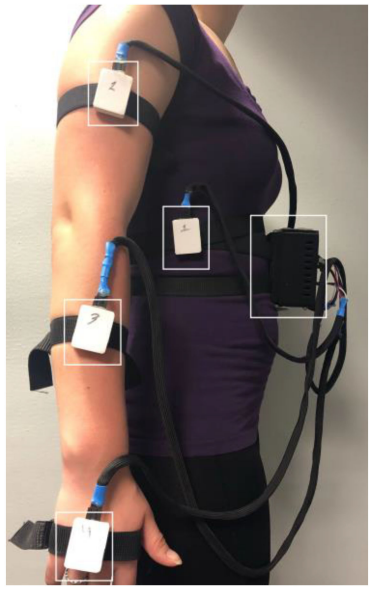

Fig. 1. Device placement on participant. Sensors were placed on the lateral aspect of the torso, upper arm, forearm, and hand. Sensors were placed midway on each body segment. The Raspberry Pi and battery were centered on the front of the abdomen.

II. Methods

A. Study Sample

50 participants were primarily recruited from the local Champaign-Urbana region using electronic advertisements. Word-of-mouth recruitment was also used. Interested parties contacted research staff indicating interest and were given a link to complete an online intake survey, gathering basic demographics information, past upper extremity injury history, and inclusion/exclusion criteria. The inclusion criteria for the study were: 1) 18-64 years of age, 2) willing to come to Urbana-Champaign for lab testing, and 3) willing to perform light resistance training. Exclusion criteria included: 1) unable to perform upper extremity movements due to physical immobilization (casts, splints, etc.), 2) unwilling to exercise without a shirt or compression clothing, 3) diagnosed with a neurological disorder that affects motor skills, 4) unable to perform a jumping jack or throw a ball, 5) pregnant, and 6) BMI greater than 30.0. 60% of the sample were males. The mean age was 21.9 $\pm$ 4.0 years of age. The mean BMI was 22.6 $\pm$ 2.7. This study was approved by the Institutional Review Board at the University of Illinois at Urbana-Champaign (Protocol Number: 18122).

B. Testing Protocol

Upon arrival to the testing site, informed consent was obtained from participants. The IMU-based device was then secured to the participant using elastic straps (Fig. 1). Triaxial IMUs were oriented with the top of each board facing the right side of the body. Participants were asked to complete 9 exercises: 1) standing row, 2) external rotation with arm abducted 90°, 3) external rotation, 4) bicep curl, 5) forearm pronation/

supination, 6) wrist curls, 7) lateral arm raise, 8) front arm raise, 9) and horizontal abduction. Prior to each exercise, research staff taught participants how to perform each movement to ensure relatively consistent form across participants. Participants could practice until exercises could be performed without coaching. Exercises were completed using either lightweight resistance bands, a 2 lb. dumbbell, or no resistance at all.

Kinematic data were recorded for 4 seconds. Participants were asked to spend one second on the lifting phase and one second on the lowering phase with no pause at the maximum range of motion. Each exercise was repeated 10 times in single repetition trials for a total of 90 trials. Participants could rest as needed to ensure good, consistent form. Each sensor recorded orientation in the form of a quaternion. Joint angles were obtained by calculating the relative rotation between a pair of sensors. Joint angles were then converted from quaternions to Euler angles, a more interpretable measure. Each trial was segmented as the paper was part of a larger study that sought to validate the IMU-based device against a motion capture system for the purpose of developing a portable device that could be used in physical therapy settings to measure exercise quality via kinematic analysis. Each trial was collected individually to allow for kinematic analysis as the two data sets needed to be compared frame-by-frame, and continuous data would have introduced an additional layer of computation that could impact validation.

C. Data Cleaning

There were two major problems with the data collected by the device: 1) the sensors sometimes jumped from the positive axis to the negative axis (or vice versa) as the sensors were constrained to a range of values (e.g., -180 to $+180$ rather than 0 to 360 degrees), and 2) the device had an inherent deficiency with clock-stretching that resulted in sporadic incorrect data points.

To address the first problem, the jumps were identified by taking the product of consecutive datapoints. The theory behind this was that if a flip in axis occurred whether it was -180 to $+180$ or -1 to $+1$, that product would be negative. To differentiate between these two scenarios, the product of the two points must be less than $-25{,}000$ (approximately the square of 158). This would allow small, correct changes in angle while filtering large jumps. This process was applied to the entire data series, and the index of each value that satisfied the previous condition was noted. Each consecutive pair of indices marked a segment of the data series that underwent an axis flip. To correct these errors, the direction of the flip needed to be identified. If the sign of the first value of each pair was positive, this indicated that the values in the segment went from $+180$ to -180, and 360 degrees was added to all values in the segment to correct for the error. If the sign of the first value of each pair was negative, this indicated the opposite, and 360 degrees was subtracted from all values in the segment.

To address the second problem, the errors were identified by calculating change between consecutive points. If a difference was greater than or equal to 5 times the distance (time) from the last "good" data point, the current value was considered an outlier. If the difference was less than the previous quotient, then the current value was not changed, and the last "good" data point was updated with the current value. All outliers were replaced with a piecewise cubic spline interpolation using the unchanged values.

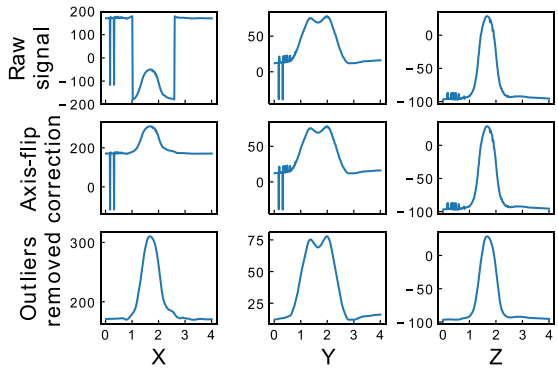

Fig. 2. Plots demonstrating the correction algorithm implemented to clean data. Raw signal included errors due to axis flipping ($+180$ to -180) and drops in data (sudden deviations from the curve). Axis-flipping was corrected by adding or subtracting 360 to flipped segments. Outliers were replaced with cubic spline interpolation. A 2nd order Savitzky-Golay filter with an 11 data points window was applied to smooth data.

After filling missing data and correcting errors, a 2nd order Savitzky-Golay filter with 11-data-point window was applied. Fig. 2 shows the cleaning steps applied to sample data.

D. Data Preparation

There was a total of 4,500 trials in the dataset. The raw dataset had an average sampling rate of 64 Hz. Simulated reduced sampling rate datasets were generated by sampling every second, third, fourth, sixth, and twelfth points from the raw datasets. Cleaning methods were applied after down sampling to prevent confounding the new datasets as down sampling could alter the waveform shape.

Two feature sets were created from the data. The first feature set consisted of a flattened structure of the data. The nine curves each were padded with 0s to match the length of the longest trial. The padded data were then flattened into a single array as a feature. The second feature set consisted of range of motion (ROM) values calculated for each of the nine curves. Range of motion was calculated for each trial by subtracting the minimum measured angle from the maximum measured angle.

E. Machine Learning Models

With raw accelerometer, magnetometer, and gyroscope data, the values are too abstract for a human to identify an exercise and call for more complex machine learning techniques. However, this study uses joint angle data which provide each exercise with a unique kinematic "fingerprint" that can easily be identified by humans. Thus, we elected to use simpler classification algorithms in this study as we believe the classification problem is not particularly complex. Using the scikit-learn library for Python, models were constructed using Random Forest (RF), LinearSVC, k-Nearest Neighbors (kNN), and Multilayer Perceptron (MLP) algorithms to classify the nine different exercises.

RF custom parameters included 300 trees with a maximum depth of 20 nodes. LinearSVC custom parameters included 200,000 max iterations and primal optimization. kNN models were constructed on the range of 1 to 10 neighbors. MLP hyper-parameters were optimized using a randomized search on hidden layer sizes, activation, solver, alpha, and learning rate.

Three analyses were performed: 1) to identify the best performing classification algorithm and feature set using a random selection of data for training and testing, 2) to evaluate the impact of different training group sizes using a random selection of data for training and testing, and 3) to evaluate the impact of different training group sizes using subject stratified training and testing.

For comparison of different models and feature sets, a 50-50 train-test split was used. Accuracy, precision, recall, F1-score, speed, and support were used to assess classifier performance. Precision is defined as the fraction of true positives relative to the total number of predicted labeled. Recall is defined as the fraction of true positives relative to the total number of true labeled. F1-score is defined as the harmonic average of precision and recall where a score of 1 represents high performance. Support is defined as the number of trials predicted for each label. Confusion matrices were also shown for some classifiers.

For comparison of different training group sizes, RF models were used. The flattened and ROM feature sets were combined into a single feature set for these analyses. Training group splits ranged from 1% to 50%. A second analysis randomly split the data by participants rather than a randomly mixed sample to assess classifier performance on unseen participants. Training group splits ranged from 10% to 98%. For each training group split, the average classifier accuracy was calculated over ten randomly blinded samples, and 10-fold cross validation was used for each sample.

III. Results

A. Feature Set Selection and Comparison of Machine Learning Algorithms

Table I shows classifier performance with RF, Linear SVC, 3-NN, and MLP algorithms applied to the flattened and ROM feature sets. Overall classifier performance was excellent with an average accuracy of 90.6%. Models using the flattened feature set consistently outperformed models using the ROM feature set (97.2% accuracy vs. 91.0% accuracy, respectively). However, models using the ROM feature set were noticeably faster to train and test.

The RF models had the greatest accuracy (98.6%) in comparison to 3-NN (97.4%) and MLP (95.7%) for the flattened feature set. The same trend was true for models using the ROM feature set. MLP models took at least two orders of magnitude longer to train in comparison to RF and 3-NN; however, the testing time for MLP was similar to RF and faster than 3-NN. RF and MLP models had longer train times and shorter test times whereas 3-NN had shorter train times and longer test times. LinearSVC was not viable for either feature set. The ROM feature set had poor accuracy (75.6%), and the flattened feature set required too much time to compute.

TABLE I
Classifier Performance Applying Random Forest (RF), K-Nearest Neighbors (3-NN), and Multilayer Perceptron (MLP) Algorithms With Flattened Kinematic and Range of Motion (ROM) Feature Sets. Accuracy, Training Time, and Test Time Are Shown for 50% Training Sets

	Accuracy	Train time (s)	Test time (s)
RF (flattened)	98.6%	12.4	.678
RF (ROM)	91.9%	1.35	0.116
3-NN (flattened)	97.4%	0.820	13.264
3-NN (ROM)	91.8%	0.003	0.063
MLP (flattened)	95.7%	1920.15	0.549
MLP (ROM)	89.3%	360.57	0.004
Linear SVC (ROM)	75.6%	0.048	0.001
Flattened average	97.2%		
ROM average	91.0%		
Total average	94.1%		

B. Exercise Specific Performance

Fig. 3 shows the confusion matrices for RF models using the flattened and ROM feature set. The RF model using the flattened feature set was 100% accurate on five exercises: 1) standing row, 2) external rotation with arm abducted 90°, 3) bicep curl, 4) wrist curl, and 5) horizontal abduction. All models using the flattened feature set had the tendency to misclassify lateral and front arm raises as each other. Models using the ROM feature set had two patterns of misclassification: 1) external rotation with the arm abducted 90° and external rotation, and 2) lateral arm raises, front arm raises, and horizontal abduction. These models consistently classified four exercises well: 1) standing row, 2) bicep curl, 3) forearm pronation/supination, and 4) wrist curls.

C. Impact of Down Sampling Data on Classifier Performance

Table II shows performance metrics of RF classifiers using the flattened feature set at six different sampling rates. All classifiers met the goal of at least 90% classification accuracy. There was no significant change in classifier performance when decreasing the sampling rate to 32 Hz (half), 22 Hz (third), or 16 Hz (quarter). Precision, recall, and F1-score values were within .01 of those of the raw classifier. Classifier performance did not begin to diminish until reducing the sampling rate to 11 Hz (sixth). Precision, recall, and F1-score decreased slightly, about 0.02 for each metric compared to the raw classifier. Further decrease in classifier performance was seen with 5 Hz data, but the classifier still met the goals. Precision, recall, and F1-score decreased by about .03 compared to the raw classifier.

Misclassification patterns persisted with down sampling data. Front and lateral arm raises continued to be misclassified as each other, and the magnitude of the errors increased as the data was further down sampled. One interesting discovery was the shape of the graphs after down samplings (Fig. 4). The overall shape of the curves remained similar for the quarter and

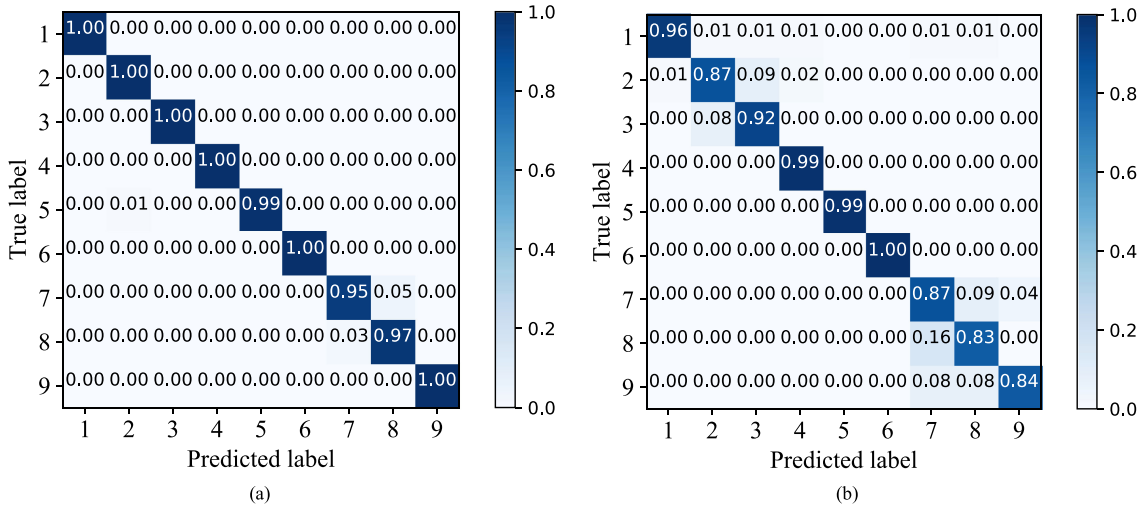

Fig. 3. Confusion matrices for random forest models using the (a) flattened and (b) ROM feature sets. Exercises are as follow – 1: standing row; 2: external rotation with arm abducted 90°; 3: external rotation; 4: bicep curl; 5: forearm pronation/supination; 6: wrist curl; 7: lateral arm raise; 8: front arm raise; 9: horizontal abduction.

TABLE II
PERFORMANCE METRICS FROM RANDOM FOREST CLASSIFICATION OF NINE UPPER LIMB EXERCISES AT THREE SAMPLING FREQUENCIES. PRECISION IS DEFINED AS THE FRACTION OF TRUE POSITIVES RELATIVE TO THE TOTAL NUMBER OF PREDICTED LABELED. RECALL IS DEFINED AS THE FRACTION OF TRUE POSITIVES RELATIVE TO THE TOTAL NUMBER OF TRUE LABELED. F1 SCORE IS DEFINED AS THE HARMONIC AVERAGE OF PRECISION AND RECALL WHERE A SCORE OF 1 REPRESENTS HIGH PERFORMANCE. SUPPORT IS DEFINED AS THE NUMBER OF TRIALS PREDICTED FOR EACH LABEL

	Precision			Recall			F1			Support		
Exercise	64 Hz	11 Hz	5 Hz	64 Hz	11 Hz	5 Hz	64 Hz	11 Hz	5 Hz	64 Hz	11 Hz	5 Hz
Standing row	0.988	0.996	0.966	1	0.94	0.97	0.994	0.967	0.968	237	249	234
Ext. rot. w/ arm abducted 90°	0.988	0.972	0.975	1	0.996	0.963	0.994	0.984	0.969	246	247	245
External rotation	0.992	0.949	0.972	0.988	0.988	0.957	0.99	0.968	0.965	242	247	256
Bicep curl	1	0.98	0.945	0.996	0.996	0.964	0.998	0.988	0.954	258	242	248
Forearm pronation	0.992	0.974	0.938	0.984	0.987	0.972	0.988	0.981	0.955	252	232	250
Wrist curls	1	1	0.963	0.996	0.971	0.943	0.998	0.985	0.953	267	244	245
Lateral arm raise	0.968	0.931	0.923	0.948	0.931	0.912	0.958	0.931	0.918	252	262	250
Front arm raise	0.948	0.914	0.906	0.963	0.933	0.945	0.956	0.923	0.923	246	252	254
Horizontal abduction	0.996	0.989	0.984	0.996	0.964	0.944	0.996	0.976	0.964	250	275	268
Macro average	0.986	0.967	0.952	0.986	0.967	0.952	0.986	0.967	0.952	2250	2250	2250

sixth down sampling in comparison to the original raw curve. Despite the classifier still performing quite well, the twelfth down sample curve looks drastically different and does not provide any clinically relevant information to patients other than the direction of movement.

D. Impact of Training Group Size on Classifier Performance

Table III shows performance metrics of RF classifiers using the combined feature set with different training group sizes ranging from 1% to 50%. Accuracy decreased about 1.5% as the training group size decreased from 50% to 30% of the sample. Accuracy further decreased to 94.1% and 91.2% with a training split of 20% and 10%, respectively. Accuracy decreased below 90% with a training split of 5% and rapidly decreased with a training split of 2% and 1%.

Fig. 5 shows performance metrics of RF classifiers using the combined feature set with participant stratified training and testing groups. With a 10% training size (5 participants for a total of 50 trials for each exercise), classifier performance was acceptable with an accuracy of 90.2%. The accuracy did not significantly change as training group size increased to 80%. Classifier performance was inconsistent with an 80% training group split as indicated by the large standard deviation. Classifier performance increased significantly with a training group split of at least 90%. 99.9% accuracy was achieved with a 98% training group split (classifying only one person).

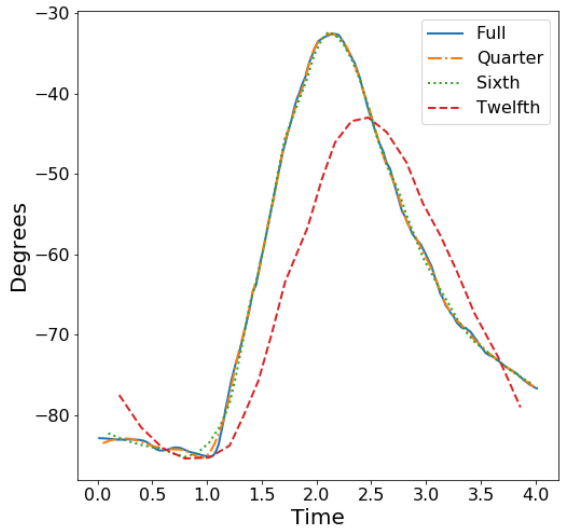

Fig. 4. Graph showing changes in kinematic data with down sampling due to applied filters and cleaning algorithm during external rotation with arm abducted 90°.

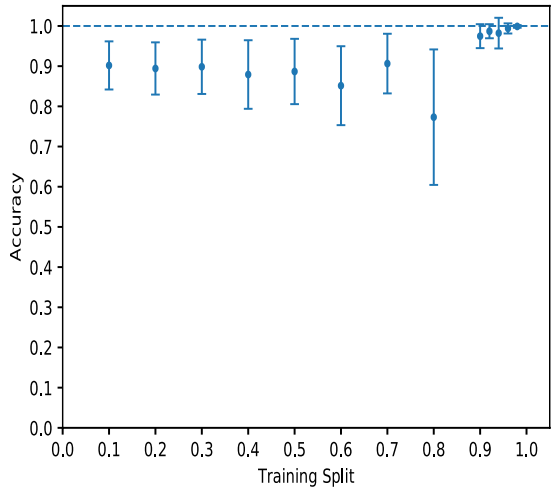

Fig. 5. Performance metrics of RF classifiers using the combined feature set with different participant stratified training group sizes ranging from 10% to 98%. Error bars indicate average standard deviation of 10 randomly blinded samples.

TABLE III
PERFORMANCE METRICS OF RF CLASSIFIERS USING THE COMBINED FEATURE SET WITH TRAINING GROUP SIZES RANGING FROM 1% TO 50%

Train split	Accuracy	Train time (s)	Test Time (s)
50%	98.6%	12.4	0.678
40%	97.3%	9.46	0.84
30%	97.1%	6.54	0.98
20%	94.1%	4.28	1.11
10%	91.2%	1.87	1.23
5%	87.2%	0.913	1.33
4%	85.6%	0.743	1.34
3%	82.0%	0.564	1.34
2%	72.7%	0.386	1.31
1%	60.1%	0.314	1.58

IV. DISCUSSION

This study found that machine learning methods could accurately classify a variety of upper extremity exercises using biomechanical data from an IMU-based device. The highest accuracy (98.6%) was attained using a random forest method–a level of accuracy surpassing the study goal of 90% and similar to the accuracies (96.85% to 97.5%) attained with other IMU devices [23], [24] and to the accuracies (95.2% to 98.3% accuracy) attained with RGB-D cameras [10], [25]. The greater performance in this study compared to previous studies could be explained by the additional IMUs allowing for additional joint kinematic data that is not ascertainable with simpler sensor designs. Each exercise can be identified with a "biomechanical fingerprint" in that there is a unique combination of movement in each plane and joint. Furthermore, previous work has shown that multi-sensor systems outperform single sensor systems when classifying continuous data and do not require complex classification algorithms which may improve processing speed and usability [26]. One study did achieve near perfect activity recognition of 10 CrossFit exercises using a wrist and ankle sensor, but repetition counting performance was poor with an average error rate of 6.1% (range: 2%–20%) [27].

The primary benefit of reducing sensors is operating under the assumption that reduction of sensors improves ergonomics and clinical usability. However, this device already uses the minimum number of IMUs necessary for kinematic analysis as each body segment requires an IMU to calculate relative joint angles. Furthermore, clothing with integrated IMUs have been developed that would make putting on multiple sensors as easy as a single sensor [28]. Fewer IMUs may also actually result in worse processing time as the data structure would shift from relative joint angles (3DOF per joint) to raw sensor data (9DOF per sensor). In order to retain the ability to classify exercises while the body is in any orientation, this would require a minimum of two sensors (18DOF): a sensor on the torso for orientation, and a sensor on the wrist or hand for assessing exercises (sensors more proximal to the torso would fail to capture distal joint exercises). Thus, this data would be approximately 50% larger than this paper's data and negatively impact processing time. Furthermore, using raw sensor data requires more consistent placement of sensors as accelerations can differ greatly depending upon placement along a limb. With relative joint angle data, the sensor placement is more resistant to inconsistent sensor placement.

Classifier performance was consistent regardless of the joint(s). Furthermore, the classifier performed similarly for the one compound exercise (standing row) as it did for the isolation exercises. The classifier's lack of specificity for certain exercises shows promise that the classifier should perform well with additional exercises. Unexpectedly, the classifier struggled with exercises that are similar in motion but performed in different planes as shown by the higher rates of misclassification for front and lateral arm raises. Similar patterns of misclassification

between abduction and forward elevation and an even stronger pattern (up to 21% of trials misclassified) between internal and external rotation have been found in another study [23]. However, the previous study used only a single IMU. This study should have been capable of discerning similar movements in different planes as relative 3D joint angles were calculated between pairs of IMUs. It is possible that by using nearest relative rotations instead of a specific rotation order could lead to similar angular data. In validating IMU-based motion capture, greatest reliability is achieved by using different rotation orders to calculate humerothoracic joint angles depending upon the arm position [29]. More accurate classification could also be attained by retaining individual sensor orientations in addition to the joint angle calculations. This could help differentiate between forward and lateral motions. Furthermore, alternative pairs of IMUs may be needed to more accurately calculate transverse movements. For example, the forearm may be a better measure of internal and external rotation compared to upper arm as skin artifacts cause IMUs on the upper arm to move disproportionately to the humerus.

With some machine learning methods, high accuracy was attainable with train and test times that would be reasonable in a clinical setting. RF and k-NN models performed similarly in this study, but the differences in training and testing time can impact device usability. If real time activity classification is desired, testing time is a priority as patients will not use a device that significantly hinders exercise. RF algorithms offer a balance between classification accuracy and speed. If additional speed is necessary, the ROM feature set may be used at the cost of accuracy. However, the previously mentioned suggestions of improving accuracy may make up for this loss.

Deep learning algorithms may also be applied to increase accuracy compared to simpler RF or k-NN algorithms [23], [30]. However, deep learning algorithms require more time to train models, which can impact usability and device adoption. A new method of learning features called probabilistic First-Take-All could be applied to accelerate the training and testing speeds in deep learning models with marginal changes to accuracy [30]. Notably, the sample size of this study (i.e. classifying 2250 exercise movements) is much larger than in a clinical setting, where the task is to classify one exercise movement at a time.

The classifier's performance did not deteriorate substantially with down-sampled data. The classifier's excellent performance despite using inaccurate down sampled data supports previous findings, where inaccurate sensors have limited impact on classification [31]. Having an IMU device collect data at a lower sampling rate has several benefits. An important benefit is longer battery—a priority in clinical settings [32]. Second, with fewer data points, the classifier can be trained and tested significantly faster, thus increasing the feasibility of real time processing of IMU data. Lower sampling frequencies may also allow the Pi to keep up with data calls – decreasing the number of random outliers – and hence decrease or eliminate the need for intensive cleaning and filtering algorithms. However, if cleaning and filtering are still required, the kinematic data may no longer be representative of the real movement and may not be useable in more detailed classifiers, such as classifying proper versus improper form.

A valid concern that many classification studies face is the generalizability of the findings. This study attempted to address this through two approaches: 1) under the premise that patients would build their own training data, and 2) under the premise that patients would utilize a previously created database as training data. The decrease in performance as the training split decreased from 50% to 10% suggests patients might not need large amounts of training data for the classifier to perform well. While a 10% split in this study consists of 50 trials per exercise, a seemingly large number, a single session of PT typically includes at least 30 repetitions per exercise. One advantage for this approach is that the classifier can be tuned to the patient's current functional status. For example, patients who are beginning PT may perform exercises with limited range of motion, which can impact classification. Furthermore, this allows for good generalizability as the device's use will not be restricted to exercises in a database, but any exercises may be performed as long as a sufficiently large training set is created.

The findings show that classifier performance can be dependent upon using previously seen data. Classifier performance decreased nearly 10% when testing on data from unseen participants. Similar decreases have been observed in other studies [23]. Using a pre-existing database to classify exercises may require many trials as accuracy did not exceed 90% until a 70% training split was used. Overfit is also possible as the consistency of classifier performance decreased as the training split increased, particularly for the 80% training split, which also saw a large drop in accuracy. This may be due to large between-subject variation. However, a large enough dataset may contain enough information to handle between-subject variation as shown by the high accuracy and low standard deviation when using training splits greater than 90%.

There were several study limitations. First, the study data set consisted of segmented trials and does not address the accuracy possible with data from a continuous exercise session, as has been done in other studies [23]. In practice, exercise data will not be segmented but must be extracted from a continuous strip of movement data. This increases the difficulty of classification as exercises need to be separated from other random movement patterns such as walking prior to exercise classification. However, the use of multiple sensors allows for improved filtering of random movements as recognized exercises will have a very characteristic movement pattern. Previous work has shown 94-96% accuracy is possible when classifying continuous data with multiple data inputs [26], [33].

Second, the study results regarding down-sampled data may not apply to classifying exercises that are more complicated, dynamic, or rapidly performed than those of this study. For example, such exercises can be prescribed in "functional PT", which uses and sport-specific movements. 90% classification accuracy was obtained when classifying movements designed to challenge mobility and stability despite using kinematic data collected with a motion capture system [34]. Third, the study evaluated only a few of the many machine learning methods.

Fourth, as discussed above, there are concerns about the generalizability of the specific classifiers developed in this study. While the study included exercises involving all three joints of the upper extremity and movements occurred in all planes of motion, classifier accuracy could be less when applied to exercises not included in the study.

Finally, the study sample was relatively homogeneous and consisted of healthy, young adults. In obese adults, subcutaneous fat may increase slippage in IMU location or increase movement artifact. Participants with impairments may also have inconsistent exercise performance resulting in decreased accuracy. However, since our data structure consists of joint kinematic data rather than raw sensor data, our classifiers should perform well even on participants with impairments under the assumption that the participants with impairments are still able to perform exercises with proper form. That is, a properly performed exercise will look the same regardless of who is performing the exercise. Factors such as length of an exercise repetition or movement speed may lower classifier performance. However, the participants in this study were given limited coaching resulting in a wide variety of performance despite an overall healthy sample. Participants were only told to perform exercises over their complete range of motion, spend one second on the lifting phase, spend one second on the lowering phase, and to not pause at maximum range of motion. Trials were only repeated if there was an issue with data collection (device failure or motion tracking issues) or if participants were performing exercises incorrectly (e.g., movement in the wrist during a bicep curl or abducting the arm during external rotation). This allowed for greater variability in the kinematic data.

In order to begin clinical testing, additional studies need to be performed: 1) repeat validation of the device on healthy and impaired participants but with continuous data streams to simulate real-world use, 2) usability analysis with a revised device that incorporates feedback from this study (e.g., wireless sensors, integrating the sensors into clothing, reducing the footprint of the Raspberry Pi hub), 3) repeat exercise classification analyses with the continuous data. The major challenge is segmenting data. In participants with impairments, it may become more difficult to identify exercise patterns if there are additional movements at the start or finish of exercise. Furthermore, with different physical impairments, these additional movement patterns may not be consistent and depend upon the impairment. Aggressive filtering and thresholds could be used to remove these patterns prior to using a sliding window approach to identify timepoints. The timepoints can then be used on the original data to ensure the actual movement pattern is intact. Previous work has shown activity recognition is possible with reasonable accuracy (82.6%–88.8%) in children and adolescents with cerebral palsy [35]. Once exercise patterns have been segmented, we would not expect a change in classifier performance as kinematics are independent of the performer.

V. Conclusion

Upper extremity PT exercises can be classified with a high degree of accuracy using kinematic data captured by IMUs. This sets forth the basis for clinical testing to better assess at-home PT. Future devices should use a Random Forest classifier and flattened data as features, as these models had the greatest accuracy with an acceptable test time. However, contrary to the trend seen in exercise wearables, high sampling frequencies are not needed for the purposes of machine learning, and sampling rates ranging from 10 Hz to 30 Hz should suffice. Lower sampling frequencies will result in faster data processing and exercise classification as well as improve device battery life, an essential consideration to wearables. It would be best for participants to generate their own training data under the supervision of their physical therapist. The main reason for this is that patients are often limited in unique manners, and a pre-existing dataset might exclude a user's exercise kinematic patterns despite being correct with respect to the user's abilities and limitations. While our sample consisted of healthy adults, we showed that it is possible to achieve perfect classification if the dataset is large enough to accommodate a wide variation in exercise performance. One possible solution is an open-access database where researchers may deposit data, though variations in data format (sampling rate, hardware, etc.) may pose a challenge, which would require standardization.

Clinical studies should explore the two approaches to training classifiers for real world use – generating patient-specific training data versus using a general pre-existing database of training data. Specifically, PT patients may fatigue easier and have larger variation between repetitions, which may benefit more from the former approach. It is possible that participants with impairments will be unable to perform exercises to the same degree as healthy participants and thus cannot be accurately classified. In this scenario, training data can be created from therapist-supervised exercise sessions resulting in a new attainable "correct" kinematic pattern.

Device usability should also be evaluated. The current device is cumbersome and difficult to use, and ergonomic improvements, such as wireless sensors and reducing the device footprint, should be made to reduce burden on the user. Patients, particularly those with physical limitations, will likely be inconsistent in sensor placement, and the impact of sensor placement on classification accuracy should be examined.

Future work should focus on clinical testing of IMU wearables to study and augment PT. With objective records of at-home PT exercise, researchers and therapists will have more insight into why PT is sometimes ineffective and be able to develop more effective solutions. Rehabilitation protocols are ideally personalized for each patient according to their recovery timeline, but it requires therapists to assess and monitor patient progress at a detailed level, which can be difficult due to time limitations. Automated exercise analysis through kinematic analysis and exercise classification will deliver additional information to patients, and therapists will be able to identify which aspects of a protocol may need additional focus despite not actually supervising the exercise. Implementation of multi-category machine learning models to identify specific errors in technique will further help patients understand how to improve their exercise when unsupervised. Additional sensors may be integrated with IMUs in the future to improve performance. Electromyography

can identify which muscles are being activated to better assess the effectiveness of exercises in rehabilitation and may improve classifier models through additional features such as muscle fatigue [36]. Thermal sensors can improve relative orientation measures and have been shown to significantly improve classification accuracy from 75 to 94% [37].

Current remote physical therapy solutions with video calls are ineffective as a single camera perspective over video call is not enough to accurately assess performance. The integration of these wearable sensors to the IoT, would allow for objective and personalized feedback to be provided to end users via a smartphone or smartwatch, and physical therapists may be able to monitor patient progress remotely. Physical therapists would be able to see exactly how patients perform exercises in all dimensions.

IMU-based kinematic analysis can guide patients through rehabilitation by reinforcing proper technique during unsupervised exercise. IMUs also allow for measuring metrics that were previously difficult to obtain such as range of motion and adherence to rehabilitation protocols. Combined metrics of exercise quality and quantity can ultimately be used to improve exercise self-efficacy and rehabilitation outcomes.

References

[1] S. M. Anwer, A. P. Alghadir, and J.-M. P. Brismee, "Effect of home exercise program in patients with knee osteoarthritis: A systematic review and meta-analysis. [Review]," *J. Geriatr. Phys. Ther.*, vol. 39, no. 1, pp. 38–48, 2015, doi: 10.1519/JPT.0000000000000045.

[2] S. J. Gilmore, J. A. McClelland, and M. Davidson, "Physiotherapeutic interventions before and after surgery for degenerative lumbar conditions: A systematic review," *Physiotherapy*, vol. 101, no. 2, pp. 111–118, Aug. 2014, doi: 10.1016/j.physio.2014.06.007.

[3] R. Zätterström, T. Fridén, A. Lindstrand, and U. Moritz, "Rehabilitation following acute anterior cruciate ligament injuries – a 12-month follow-up of a randomized clinical trial," *Scand. J. Med. Sci. Sports*, vol. 10, no. 3, pp. 156–163, Jun. 2000.

[4] M. F. Pisters, C. Veenhof, F. G. Schellevis, J. W. R. Twisk, J. Dekker, and D. H. De Bakker, "Exercise adherence improving long-term patient outcome in patients with osteoarthritis of the hip and/or knee," *Arthritis Care Res.*, vol. 62, no. 8, pp. 1087–1094, Aug. 2010.

[5] S. F. Bassett, "The assessment of patient adherence to physiotherapy rehabilitation," *N. Z. J. Physiother.*, vol. 31, no. 2, pp. 60–66, 2003.

[6] G. S. Kolt and J. F. McEvoy, "Adherence to rehabilitation in patients with low back pain," *Man. Ther.*, vol. 8, no. 2, pp. 110–116, May 2003.

[7] G. L. Moseley, "Do training diaries affect and reflect adherence to home programs?" *Arthritis Care Res.*, vol. 55, no. 4, pp. 662–664, Aug. 2006.

[8] S. Hong and Y. Kim, "Dynamic pose estimation using multiple RGB-D cameras," *Sensors*, vol. 18, no. 11, Nov. 2018, Art. no. 3865.

[9] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Comput. Vis. Image Underst.*, vol. 171, pp. 118–139, Jun. 2018.

[10] I. Ar and Y. S. Akgul, "A computerized recognition system for the home-based physiotherapy exercises using an RGBD camera," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 6, pp. 1160–1171, Nov. 2014.

[11] C. Kertesz, "Physiotherapy exercises recognition based on RGB-D human skeleton models," in *Proc. Eur. Modelling Symp.*, Nov. 2013, pp. 21–29.

[12] R. Komatireddy, "Quality and quantity of rehabilitation exercises delivered by A 3-D motion controlled camera: A pilot study," *Int. J. Phys. Med. Rehabil.*, vol. 02, no. 04, 2014.

[13] M. Al-Amri, K. Nicholas, K. Button, V. Sparkes, L. Sheeran, and J. Davies, "Inertial measurement units for clinical movement analysis: Reliability and concurrent validity," *Sensors*, vol. 18, no. 3, p. 719, Feb. 2018.

[14] D. Dinu, M. Fayolas, M. Jacquet, E. Leguy, J. Slavinski, and N. Houel, "Accuracy of postural human-motion tracking using miniature inertial sensors," *Procedia Eng.*, vol. 147, pp. 655—658, 2016.

[15] V. Joukov, M. Karg, and D. Kulic, "Online tracking of the lower body joint angles using IMUs for gait rehabilitation," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Medicine Biol. Soc.*, Aug. 2014, pp. 2310–2313.

[16] D. Whelan, M. O'Reilly, B. Huang, O. Giggins, T. Kechadi, and B. Caulfield, "Leveraging IMU data for accurate exercise performance classification and musculoskeletal injury risk screening," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Medicine Biol. Soc.*, Aug. 2016, pp. 659–662.

[17] W. Xu, C. Ortega-Sanchez, and I. Murray, "Measuring human joint movement with IMUs: Implementation in custom-made low cost wireless sensors," in *Proc. IEEE 15th Student Conf. Res. Develop. Putrajaya*, Dec. 2017, pp. 172–177.

[18] M. A. O'Reilly, D. F. Whelan, T. E. Ward, E. Delahunt, and B. M. Caulfield, "Classification of deadlift biomechanics with wearable inertial measurement units," *J. Biomech.*, vol. 58, pp. 155–161, Jun. 2017.

[19] M. A. O'Reilly, D. F. Whelan, T. E. Ward, E. Delahunt, and B. M. Caulfield, "Technology in strength and conditioning: Assessing bodyweight squat technique with wearable sensors," *J. Strength Cond. Res.*, vol. 31, no. 8, pp. 2303–2312, Aug. 2017.

[20] R. Argent, P. Slevin, A. Bevilacqua, M. Neligan, A. Daly, and B. Caulfield, "Wearable sensor-based exercise biofeedback for orthopaedic rehabilitation: A mixed methods user evaluation of a prototype system," *Sensors*, vol. 19, no. 2, Jan. 2019.

[21] N. D. Harada, S. Dhanani, M. Elrod, T. Hahn, L. Kleinman, and M. Fang, "Feasibility study of home telerehabilitation for physically inactive veterans," *J. Rehabil. Res. Dev.*, vol. 47, no. 5, pp. 465–475, 2010.

[22] L. De Baets, R. van der Straaten, T. Matheve, and A. Timmermans, "Shoulder assessment according to the international classification of functioning by means of inertial sensor technologies: A systematic review," *Gait Posture*, vol. 57, pp. 278–294, Sep. 2017.

[23] D. M. Burns, N. Leung, M. Hardisty, C. M. Whyne, P. Henry, and S. McLachlin, "Shoulder physiotherapy exercise recognition: Machine learning the inertial signals from a smartwatch," *Physiol. Meas.*, vol. 39, no. 7, 2018, Art. no. 075007.

[24] J.-I. Pan, H.-W. Chung, and J.-J. Huang, "Intelligent shoulder joint home-based self-rehabilitation monitoring system," *Int. J. Smart Home*, vol. 7, no. 5, pp. 395–404, Sep. 2013.

[25] D. Antón, A. Goñi, and A. Illarramendi, "Exercise recognition for Kinect-based telerehabilitation," *Methods Inf. Med.*, vol. 54, no. 2, pp. 145–155, 2015.

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

[27] A. Soro, G. Brunner, S. Tanner, and R. Wattenhofer, "Recognition and repetition counting for complex physical exercises with deep learning," *Sensors*, vol. 19, no. 3, p. 714, Feb. 2019.

[28] S.-W. Kang *et al.*, "The development of an IMU integrated clothes for postural monitoring using conductive yarn and interconnecting technology," *Sensors*, vol. 17, no. 11, p. 2560,, Nov. 2017.

[29] J. López-Pascual, M. L. Cáceres, H. De Rosario, and Á. Page, "The reliability of humerothoracic angles during arm elevation depends on the representation of rotations," *J. Biomech.*, vol. 49, no. 3, pp. 502–506, Feb. 2016.

[30] J. Ye, G. Qi, N. Zhuang, H. Hu, and K. A. Hua, "Learning compact features for human activity recognition via probabilistic first-take-all," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 126–139, Jan. 2020.

[31] S. Han, M. Achar, S. Lee, and F. Peña-Mora, "Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring," *Vis. Eng.*, vol. 1, no. 1, p. 6, 2013.

[32] A. Krause *et al.*, "Trading off prediction accuracy and power consumption for context-aware wearable computing," in *Proc. 9th IEEE Int. Symp. Wearable Comput.*, 2005, pp. 20–26.

[33] Q. Wang, G. Kurillo, F. Ofli, and R. Bajcsy, "Unsupervised temporal segmentation of repetitive human actions based on kinematic modeling and frequency analysis," in *Proc. Int. Conf. 3D Vision*, Oct. 2015, pp. 562–570.

[34] A. L. Clouthier, G. B. Ross, and R. B. Graham, "Sensor data required for automatic recognition of athletic tasks using deep neural networks," *Front. Bioeng. Biotechnol.*, vol. 7, p. 473, Jan. 2020, doi: 10.3389/fbioe.2019.00473.

[35] M. Ahmadi, M. O'Neil, M. Fragala-Pinkham, N. Lennon, and S. Trost, "Machine learning algorithms for activity recognition in ambulant children and adolescents with cerebral palsy," *J. NeuroEng. Rehabil.*, vol. 15, p. 105, Nov. 2018, doi: 10.1186/s12984-018-0456-x.

[36] S. F. del Toro, S. Santos-Cuadros, E. Olmeda, C. Álvarez-Caldas, V. Díaz, and J. L. San Román, "Is the use of a low-cost sEMG sensor valid to measure muscle fatigue?" *Sensors*, vol. 19, no. 14, p. 3204, Jul. 2019, doi: 10.3390/s19143204.

[37] J. Lui and C. Menon, "Would a thermal sensor improve arm motion classification accuracy of a single wrist-mounted inertial device?" *Biomed. Eng. OnLine*, vol. 18, p. 53, May 2019, doi: 10.1186/s12938-019-0677-7.