2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Sensor-Based Evaluation of Physical Therapy Exercises*

Andrew S. Whitford[1], Emily Kim[1], Eni Halilaj[2], Keelan Enseki[3], Adam Popchak[4], and Jessica Hodgins[1]

Abstract— **Physical therapy is important for the treatment and prevention of musculoskeletal injuries, as well as recovery from surgery. In this paper, we explore techniques for automatically determining whether an exercise was performed correctly or not, based on camera images and wearable sensors. Classifiers were tested on data collected from 30 patients during normally-scheduled physical therapy appointments. We considered two lower limb exercises, and asked how well classifiers could generalize to the assessment of individuals for whom no prior data were available. We found that our classifiers performed well relative to several metrics (mean accuracy: 0.76, specificity: 0.90), but often returned low sensitivity (mean: 0.34). For one of the two exercises considered, these classifiers compared favorably with human performance.**

Clinical relevance— **This work establishes a baseline level of performance for automatic classification of exercise performance in a patient population, based on two cameras or body-worn IMUs.**

I. INTRODUCTION

In recent years, there has been growing interest in automated assessment of physical therapy exercises. Physical therapy is effective in treating and preventing mobility limiting conditions and plays an important role in determining outcomes [1]. Automated and technologically-enabled assessments of exercise performance are attractive because they yield data required to track improvement, monitor outcomes, and they facilitate patients' access to expert knowledge, which can be critical for overcoming geographic and socioeconomic barriers to care [2], [3].

In this paper, we assess the capability of machine learning algorithms to detect errors in the performance of exercises commonly used in physical therapy and post-surgical rehabilitation. We recorded data from patients being seen in a clinic for conditions of the knee or hip – such as anterior cruciate ligament (ACL) injuries, osteoarthritis, or joint replacement – as well as healthy controls. The data recorded from each subject includes video from two cameras placed 90° apart and a set of 10 body-worn inertial measurement units (IMUs).

For the purpose of assessing performance, we consider two application scenarios. First, we assess generalization error on held-out exercise repetitions, which corresponds to the scenario in which automated assessment tools are trained during an in-person clinic visit, and then sent home with the patient. Second, we assess generalization error on held-out patient data, which corresponds to the scenario in which automated assessment tools are applied to patient data without any prior exposure to the individual being tested. The latter scenario is more general, but also more challenging. We find that the average accuracy, sensitivity, and specificity scores of our best classifier all exceed 0.70 for two different exercises in the first scenario, but for only one side of one exercise in the second. Finally, we compare the performance of these algorithms to metrics of human performance in labeling exercise errors from video, finding that the classifiers compared favorably for one of two exercises.

II. RELATED WORK

Prior work has shown that classifiers can detect specific patterns of movements or errors, using only data obtained from wearable sensors [4] and cameras. On a small sample set, Taylor et al. showed that classifiers trained on body-worn IMU data could detect errors with high accuracy, when the classifier is trained on data taken from the same individual that it is tested on. However, testing on unseen subjects caused classifier performance to be sharply reduced [5]. The same authors later extended these results to multi-label classification – which detects and labels multiple potential errors per exercise – targeting patients undergoing rehabilitation for knee osteoarthritis [6]. In similar work, Whelan et al. showed that random forest classifiers could detect aberrant performance of a single leg squat exercise – using IMU data from 83 healthy participants – with greater than 75% accuracy, sensitivity, and specificity [7]. However, this result was limited to a single exercise on a single side (left), used labels from a single expert, and did not test the classifier with data from unseen subjects. The same group later expanded these results – testing with unseen subjects for the purpose of deadlift exercise assessment [8] – and reported an accuracy of 0.73. Although these classifiers were reasonably sensitive (0.78) to errors, the reported specificity was low (0.49), resembling the earlier results of [5]. Similarly poor results were obtained for barbell squats (accuracy: 0.64, sensitivity: 0.70, specificity: 0.28) [9]. For bodyweight squats and lunges, results were more promising, with both accuracy and specificity exceeding 0.90 in both cases, and sensitivity of 0.96 and 0.80, respectively [10][11]. Dajime et al. developed a quality assessment system for screening

*This work was supported by the University of Pittsburgh Medical Center, the National Science Foundation, and the Center for Machine Learning and Health at Carnegie Mellon University

[1] Andrew Whitford (whitford@cmu.edu), Emily Kim (ekim2@andrew.cmu.edu), and Jessica Hodgins (jkh@cmu.edu) are with the Robotics Institute, Carnegie Mellon University, Pittburgh, PA 15213, USA

[2] Eni Halilaj is with Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA ehalilaj@andrew.cmu.edu

[3] Keelan Enseki is with the Centers for Rehab Services, University of Pittsburgh Medical Center, Pittburgh, PA 15203, USA

[4] Adam Popchak is with the Department of Physical Therapy, University of Pittsburgh, Pittburgh, PA 15260, USA

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

squat, forward lunge, and single leg squat [12] exercises. The average accuracy, sensitivity and specificity ranged from 0.74-0.85, 0.66-0.89, and 0.58-0.88, respectively.

A useful benchmark in the evaluation of automated algorithms is comparison with human performance on comparable tasks. Whelan et al. measured the intra- and inter-rater agreement – quantified via Kappa scores – within a pool of 47 physical therapists and physical therapy students asked to evaluate video-recorded exercises for errors [13]. All subjects were healthy volunteers. Overall, they found minimal-to-moderate intra-rater agreement, and minimal-to-weak inter-rater agreement. Moreover, agreement was lower for errors that occurred naturally, rather than being intentionally induced. Agreement varied by the type of exercise. These results suggest that the threshold for exceeding human performance is relatively low.

III. DATA COLLECTION

As part of a normal clinical visit, video and sensor data were collected from 30 patients performing lower limb exercises while undergoing physical therapy for conditions of the hip or knee. Data were collected during normal appointments within a sports medicine and rehabilitation facility. Fifteen common lower limb exercises were chosen for monitoring, but results for only two are reported here. Patients were eligible for participation if they would normally be asked to complete any of these exercises as part of their prescribed treatment program, were aged 18 to 85, and if their body mass index (BMI) was less than 35. Further, patients with BMI over 30 were excluded if any cardiovascular comorbities were indicated in the medical record. Data were also collected from 10 healthy controls. Select controls (7 of 10) were instructed to simulate specific errors on a subset of exercise repetitions, in order to increase the size of the training data set. All procedures were approved by the Carnegie Mellon University Institutional Review Board (IRB), as well as the University of Pittsburgh IRB, before any data were collected.

After obtaining consent, ten inertial measurement units (IMUs) were affixed to the arms, legs, feet, and torso of each subject, as indicated in Figure 1. Participants were shown video demonstrations of the relevant exercises, and asked to perform repetitions of these exercises in full view of two cameras. For upright exercises, participants stood facing camera A head-on, while camera B recorded from the participant's right side. The angle between the cameras was slightly less than 90^o. For seated exercises, or exercises done while lying down, the side-view camera was aimed at the length of an exercise bed and the front-view camera was aimed at one end of the bed.

Each Xsens Awinda (Xsens Technologies BV, Enschede, Netherlands) wearable sensor contains an accelerometer, a gyroscope, a magnetometer, a barometer, and a thermometer [14]. Onboard processing enables the sensors to compress high sample rate estimates of time-varying acceleration and orientation, and to wirelessly transmit those estimates to a data logger at 40 Hz. For this analysis, we considered the 3D acceleration and orientation estimates (expressed as quaternions) for 70 channels of IMU sensor data.

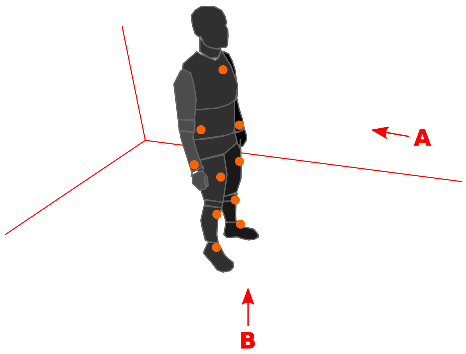

Fig. 1. Diagram of the data collection setup. Data were collected in a roughly 10 ft by 16 ft exercise space. Subjects wore a total of ten IMU sensors (orange), located on the trunk, wrists, and legs. During standing exercises, subjects faced camera A, and camera B captured video from the right side, at an oblique angle. During exercises that required the subject to lie down, an exercise bed was placed opposite camera B.

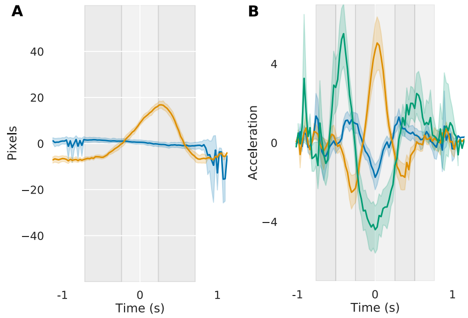

Fig. 2. Sample data. Each line represents the mean of a sensor channel across all reps for a particular exercise and side, and shaded regions indicate confidence intervals. Data are aligned to the point manually labeled as the middle of each exercise. Shaded grey bars indicate two of the feature window epochs used to generate features. The complete list of feature windows epochs can be found in Table I. A: Right knee pose keypoint measured via the side camera during the left-sided step-up exercise. B: Foot IMU accelerometer worn during the right-sided sidelying hip abduction exercise.

Color and grayscale video frames were recorded by a pair of Realsense D435 cameras (Intel Corporation, Santa Clara, CA USA) at resolutions of 1920x1080 and 840x480, respectively. A depth data stream was also recorded, but is not considered here. The variable frame rate of the cameras averaged roughly 30 Hz.

Both the IMU and video data were continuously streamed to a solid state drive (SSD) for the duration of the participant's physical therapy session.

Pose information was extracted from video data using the AlphaPose algorithm and software [15]. AlphaPose is a modern computer vision algorithm for markerless extraction of human pose information from unstructured video streams. The Resnet152 AlphaPose model pretrained on COCO dataset was individually applied to each frame of the infrared video stream. Temporal information was not considered during this pose estimation step.

The pose algorithm yields 2D pose data and confidence estimates at the frame rate of the video streams. In total, this

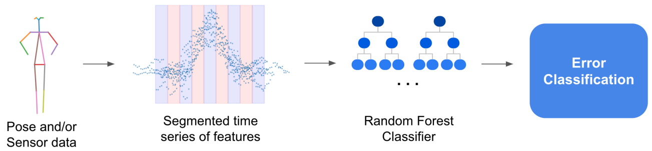

Fig. 3. Analysis pipeline. Pose and/or sensor data is segmented, and for each segment, statistics are calculated to be appended into the features. Then Random Forest classifiers are created from the features.

yields 100 channels of pose data, sampled at approximately 30Hz. The pelvis keypoint was treated as the origin, for the purpose of establishing a local coordinate system. A sample of pose data is shown in Figure 2-A.

Pose data were synchronized – between cameras, and with wearable sensor data – via an infrared signal: a pair of infrared illuminators were triggered by the XSens system at the start of data collection. The pulses emitted by the illuminators were visible in the infrared stream of each camera. All data streams were aligned to this event.

Once aligned, sensor and pose data were interpolated (cubic) and resampled with a constant 25 Hz sampling rate. This caused all data sources to share a common and consistent time base.

Once the data had been captured and processed, errors in each participant's exercise performance were evaluated. Using the color video recordings, two licensed clinicians independently labeled the timestamp of the movement errors made by each participant. Before labeling, the clinicians agreed on a small set of common errors associated with each exercise (Table II). The clinicians then reviewed any discrepancies between the two sets of labels, and agreed on a consensus set. Both the original annotations and the consensus set are considered in our analysis. The annotation procedure was the same for both patients and healthy controls.

A two-step procedure was used to segment exercise repetitions. First, the start, peak, and end of each exercise repetition were estimated from the accelerometer data. For each exercise, the peak was defined to be an easily-identified movement event that occurred mid-way through a repetition. During a second pass over the data, segmentation of repetitions was visually verified and/or manually adjusted using the video streams.

Each error annotation was associated with the closest exercise repetition. If an error timestamp fell within the temporal bounds of a repetition, then it was associated with that repetition. If an error label was assigned to a timestamp between the bounds of two repetitions, then the timestamp was moved to within the nearest bound.

IV. ANALYSIS

The objective of our analysis is to develop an algorithm capable of detecting the presence or absence of specific errors during repetitions of an exercise (Fig. 3). We framed the problem in terms of binary classification.

We chose to use a random forest classifier for the error detection analysis. Tree-based algorithms are robust to mixed types, outliers, and missing data. They are relatively fast to construct, and scale effectively. Implicit feature selection and robustness to irrelevant predictors make them well-suited to exploratory analyses, such as this. They have also been cited as a preferred method in related work [16]. With ample data and labels, modern deep learning methods might produce better prediction outcomes, but we judged a tree-based approach to be better-suited to this relatively small data set.

All analyses were implemented in Python, and the classifier was implemented using the scikit-learn package. Each random forest consisted of 400 trees. The maximum number of features to consider when looking for the best split was set to the square-root of the number of features. The remaining hyperparameters were set to the implementation default values.

Features were computed over temporal windows of varying duration and displacement. Displacement determines the center of the window relative to the peak event of each exercise repetition. An example of a set of feature windows is shown in Figure 2, where the gray bars represent windows with centers spread evenly across the exercise repetition epoch. All feature window durations and displacements are listed in Table I. The total number of feature windows per repetition was 47. For each feature window, a set of six statistics were computed: the sum, mean, median, maximum, minimum, and variance of the data within the window.

TABLE I
FEATURE WINDOW DURATIONS AND DISPLACEMENTS. ALL DISPLACEMENTS ARE SYMMETRIC ABOUT ZERO, BUT NEGATIVE VALUES ARE OMITTED HERE. THE TOTAL NUMBER OF DURATION AND DISPLACEMENT PAIRS IS 47.

Duration (ms)	Displacements (ms)
160	80, 240, 400, 560, 720, 880, 1040, 1200
280	0, 280, 560, 840, 1120
400	200, 600, 1000
520	0, 520, 1040
640	320, 960
760	0, 760
880	440
1480	0
1960	0

In order to assess how well the algorithm generalizes to new data, we considered two cross-validation (CV) approaches. In both cases, data from both patients and healthy controls were used to train the classifiers, but only data from patients were used for testing performance.

First, we considered what will be referred to as random subset cross-validation (RSS-CV). This approach is a standard way to assess generalization error in machine learning, and entails randomly selecting a fraction (30%, in this analysis) of samples (i.e., exercise repetitions) to be held-out for testing.

Second, we considered what will be referred to as leave-N-subjects-out cross-validation (LNSO-CV). Initially, we used leave-one-subject-out cross-validation (LOSO-CV), but this frequently yielded test data sets that contained only a single class (e.g., no errors). The LNSO-CV approach holds out all samples for N subjects, to be used as testing data. We consider it an important measure of how we might expect our algorithm to perform with new patients, for whom no prior data are available [17]. In this analysis, we chose groups of N=5 participants.

In each case, the mean performance metrics across 300 cross validation splits are reported. These metrics were computed using held-out, testing data from patients only. Controls were used only for training.

To facilitate comparison with human performance and prior work, we compute Kappa scores for the agreement between ground truth labels and classifier output. When physical therapists were asked to re-evaluate videos of exercise performance by healthy subjects, at least 30 days after a prior evaluation of the same recording, Whelan et al. observed only minimal to moderate agreement between the pairs of assessments [13]. For assessments in which errors occurred naturally (i.e., they were not induced), they reported aggregate Cohen's Kappa values with a mean of 0.38, and 95% confidence intervals ranging from 0.32 to 0.44. When computed individually, for each of the three exercises considered, the mean Kappa scores were 0.39, 0.40, and 0.49.

We compare classifier performance with intra-rater agreement using Cohen's Kappa [13]. This analysis treats classifier prediction output like a second set of labels from the same expert, but taken on a different occasion. We view this as equivalent to the assumption that the algorithm aims to capture the knowledge of the clinicians, in order to replicate their clinical decisions.

V. RESULTS

We selected data sets from 30 patients and 10 healthy controls for analysis. Participant ages skewed young, with 17 of 30 patients under 30 years of age. Of those remaining, only four were over 60. Among the controls, all but one were under 30. Among patients, 17 were female and 13 were male. The controls were less balanced, with eight females out of ten.

The most common condition treated was sprain of anterior cruciate ligament (nine patients), followed by pain in the knee (five patients). A total of 21 patients were treated for conditions involving the knee, and six were treated for conditions involving the hip.

After the video recordings had been annotated, we selected the two exercises with the largest number of errors for analysis. Since the overall error rate in our data set tended to be low, we elected to initially focus on those exercises that we expected could furnish enough labeled samples – of each class – to train an effective classifier.

The selected exercises were the *sidelying (S/L) hip abduction* and the *step-up*. The SL hip abduction exercise is accomplished while a subject lays on their side: the upper leg sweeps in an arc from a position parallel with the other leg, to a position in which the foot is elevated – such that the two legs form a sideways V shape. The leg is then returned to the starting position. For the step up exercise, each participant is asked to start with both feet on a small exercise block, elevated a few inches off of the ground. After stepping one foot off, to come to rest on the ground, they are asked to return to the starting position. Among the 40 participants, all completed the step-up exercise (865 total repetitions), but only 21 patients and 9 healthy control completed the sidelying hip abduction exercise (617 total repetitions). The errors for each exercise, along with the number of times each error was observed, are in Table II.

For each exercise, subsets of errors often co-occurred, and this caused some ambiguity in the labeling process. For this reason – and since the most frequent individual error accounted for only 13% and 33% of repetitions for the two exercises, respectively – the clinicians advised that we merge a subset of error types for the classification analysis. These errors are marked with a red dot in Table II. We chose to consider only these merged error classes, and to ignore the remaining (less frequent) error types. Our analysis, therefore, aims to evaluate binary classifiers that assign labels of *error* or *no-error* to each exercise repetition.

TABLE II
EXERCISES AND ERRORS

Exercise	Error	Count
Hip abduction	Hip into any amount of flexion ●	203
	Hips rolled back or forward ●	49
	Insufficient ROM	0
	Legs not straight (top leg)	34
	Motion too fast	0
Step down	Excessive foot pronation	8
	High knee valgus ●	70
	Insufficient knee flexion	40
	Knee not in line with foot ●	8
	Pelvis tilt or drop ●	110

Classifier scores for RSS-CV and LNSO-CV are summarized in Table III. A few general observations can be made about these results. First, classifiers based on features derived from IMU data always performed better than those based on features derived from video data. The combination of these two feature sets often slightly improved performance. Finally, we see a sharp decline in performance between the RSS-CV and LNSO-CV scenarios, in every single case.

For the RSS-CV analysis – in which 30% of exercise repetitions were held-out for testing – performance metrics were generally high. For the SL hip abduction exercise, every score but one was above 0.90. This is also true of the accuracy and specificity for the step-up exercise. Although they did not exceed 0.90, the MCC and sensitivity were still relatively high for this exercise. For the combined feature set (i.e., IMU+pose), all metrics exceeded 0.80. Altogether, the classifiers performed effectively.

As expected, classification of data from unseen subjects was a more challenging scenario and the results for the LNSO-CV analysis were more variable. Specificity scores were generally high, exceeding 0.85 in every instance for the combined feature data set. Accuracy scores exceeded 0.70 in every instance for the same features. However, the sensitivity scores were mostly low. The classifiers often had trouble detecting errors. That was especially true for the step up exercise. A clear exception to this poor performance is the left side of the sidelying hip abduction exercise, for which the sensitivity was around 0.70. Although these values are at least competitive with prior work [4], they do not match the best performance for exercise assessments with IMU sensor data from healthy individuals [10][11].

Although the standard errors of the mean (SEM) performance estimates were quite small in all cases (< 0.04), it is noteworthy that the variance of these estimates increased dramatically in the LNSO-CV analyses – most often having the same order of magnitude as the mean estimate. Whereas the mean MCC metric across both exercises and both sides was 0.28, for example, the standard deviation was 0.35. This indicates that the classifier could frequently perform poorly, even if performing well on average. As we discuss in the next section, this is likely attributable to a biased distribution of errors among the sampled participants.

Much of the prior work in this area sought to detect errors made by healthy subjects – often with instructions to intentionally deviate from optimal movement patterns. Such a design has the potential advantage of distributing errors among subjects. A study population with well-distributed errors helps to ensure that training data samples errors committed by a diverse set of individuals. This balance is important to the analysis of how well classifiers generalize to unseen subjects.

In this study, we collected data from patients that committed spontaneous errors, so there was no guarantee that the errors would be well-distributed among individuals. For the left side of the SL hip abduction exercise, samples for 45% (13 of 30) of participants included a mixture of exercise repetitions with and without errors. However, the percentages for the remaining three cases were much lower, at 23%. That amounts to seven of 30 participants for the right side of the abduction exercise, and nine of 40 participants for each side of the step-up exercise. We suspect that this biased distribution of errors likely affected the performance of our classifiers. Indeed, the classifiers performed quite a bit better for the left side of the SL hip abduction exercise, in which errors were distributed more evenly. A larger sample size could mitigate this effect.

Similar to the results in the previous section, the Cohen's Kappa scores for the RSS-CV analysis were high, indicating high agreement between clinician labels and the automated predictions. The mean Kappa score across all exercises and sides was roughly 0.90. That is well outside of the range reported by Whelan et al [13].

The results for the LNSO-CV analysis are again more complex, but follow the same pattern as for the general results of Table III. In aggregate – averaged across both exercises and sides – the Kappa score is quite low, at around 0.20. However, this is drawn down by the poor performance for the step-up exercise. When computed for the sidelying hip abduction exercise only, the mean Kappa score is 0.38 – equal to the mean score reported by Whelan et al. If only the left side of the abduction exercise is considered (i.e., the better-performing side), then the mean Kappa score is roughly 0.50. In at least this case, then, the classifier predictions agree with clinician labels to an extent that exceeds that of the intra-rater agreement observed among trained physical therapists. In other words, the classifier can be expected to reproduce the expert's labels better than the experts themselves can be expected to.

VI. CONCLUSIONS

Classifiers trained on data from the same subject to which they are applied can effectively detect errors in the performance of two common exercises in lower limb rehabilitation. This analysis supports the scenario in which a personalized classifier is calibrated while a patient performs exercises in a clinic.

When no prior data are available for personalization, performance of the classifiers declines substantially. This result is expected, but the magnitude of the reported decline has varied substantially in prior work. For one of the exercises we tested, performance remained reasonably high – especially on the left side. For the other exercise, classifier performance was near chance. This result might be explained by the relative magnitude of the movements involved: the exercise for which the classifier performed poorly – the step up – involved only small, subtle movements. More generally, we anticipate that classifier performance would climb with a larger sample size – primarily due to a heavy imbalance in the distribution of samples with and without errors, across the clinical population.

Classifiers using features derived from wearable IMUs consistently outperformed classifiers using features derived from video-based pose estimates. Combining the two data sources resulted in small gains relative to the IMU-only features. We suspect that the poor performance of the pose-derived features has more to do with the pose extraction method than the error classification step. In future work, we aim to improve pose extraction and explore sensor fusion.

For one of the two exercises considered, agreement between our classifiers and clinician-provided labels was comparable to the intra-rater agreement reported for physical therapists making repeated assessments. For one side of that same exercise, the classifier exceeded the reported range of human performance.

Although the performance of our classifiers did not exceed all prior work in this area, most or all of that work has dealt with healthy subjects and/or intentionally-induced errors. Our data were obtained in a physical therapy clinic. A notable consequence of collecting data from patients receiving care is that errors tend to be less frequent, and less uniformly-distributed. For example, some patients made errors on every exercise repetition, whereas others made no errors at all.

TABLE III

Accuracy, Matthew's Correlation coefficients (MCC), sensitivity, and specificity from AlphaPose keypoints (Alphapose), IMU data (sensor), and combined (both) features trained classifiers. (i) 70/30 train-test cross validation and (ii) Leave-5-out cross validation were each performed on the two exercise types, sidelying (S/L) hip abduction and step down.

		70/30 train-test cross validation				Leave-5-out-cross validation			
		S/L hip abduction		Step down		S/L hip abduction		Step down	
		Right	Left	Right	Left	Right	Left	Right	Left
accuracy	Alphapose	0.97	0.95	0.93	0.87	0.50	0.57	0.83	0.67
	IMU	0.98	0.96	0.96	0.96	0.70	0.80	0.84	0.70
	Both	0.98	0.96	0.96	0.95	0.72	0.80	0.83	0.70
mcc	Alphapose	0.94	0.89	0.60	0.63	-0.13	0.14	-0.17	-0.09
	IMU	0.96	0.92	0.83	0.89	0.39	0.58	0.14	-0.04
	Both	0.97	0.92	0.83	0.87	0.43	0.56	0.17	-0.03
sensitivity	Alphapose	0.94	0.92	0.49	0.57	0.11	0.36	0.00	0.12
	IMU	0.98	0.95	0.80	0.90	0.36	0.70	0.12	0.11
	Both	0.98	0.95	0.80	0.88	0.41	0.68	0.15	0.11
specificity	Alphapose	0.99	0.96	0.99	0.97	0.77	0.76	0.93	0.87
	IMU	0.98	0.97	0.99	0.98	0.91	0.88	0.95	0.86
	Both	0.98	0.97	0.99	0.98	0.91	0.88	0.94	0.87

This complicates cross-validation methods based on leaving subjects out, and contributes to high variability in classifier performance metric estimates. We believe that larger sample sizes will be essential for establishing high-performance automated assessments.

ACKNOWLEDGMENT

This work was made possible by funding and support from the University of Pittsburgh Medical Center, via the Pittsburgh Health Data Alliance and the Center for Machine Learning and Health. It is based upon work supported by the National Science Foundation under Grant No. IIS-1602337. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] T. Bandholm, T. W. Wainwright, and H. Kehlet, "Rehabilitation strategies for optimisation of functional recovery after major joint replacement," *Journal of experimental orthopaedics*, vol. 5, no. 1, pp. 1–4, 2018.

[2] S. K. Carter and J. A. Rizzo, "Use of outpatient physical therapy services by people with musculoskeletal conditions," *Physical Therapy*, vol. 87, no. 5, pp. 497–512, 2007.

[3] S. R. Machlin, J. Chevan, W. W. Yu, and M. W. Zodet, "Determinants of utilization and expenditures for episodes of ambulatory physical therapy among adults," *Physical Therapy*, vol. 91, no. 7, pp. 1018–1029, 2011.

[4] M. O'Reilly, B. Caulfield, T. Ward, W. Johnston, and C. Doherty, "Wearable inertial sensor systems for lower limb exercise detection and evaluation: a systematic review," *Sports Medicine*, vol. 48, no. 5, pp. 1221–1246, 2018.

[5] P. E. Taylor, G. J. Almeida, T. Kanade, and J. K. Hodgins, "Classifying human motion quality for knee osteoarthritis using accelerometers," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 339–343.

[6] P. E. Taylor, G. J. Almeida, J. K. Hodgins, and T. Kanade, "Multi-label classification for the analysis of human motion quality," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 2214–2218.

[7] D. F. Whelan, M. A. O'Reilly, T. E. Ward, E. Delahunt, and B. Caulfield, "Technology in rehabilitation: Evaluating the single leg squat exercise with wearable inertial measurement units," *Methods of Information in Medicine*, vol. 56, no. 02, pp. 88–94, 2017.

[8] 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," *Journal of Biomechanics*, vol. 58, pp. 155–161, 2017.

[9] D. F. Whelan, M. A. O'Reilly, T. E. Ward, E. Delahunt, and B. Caulfield, "Technology in rehabilitation: Comparing personalised and global classification methodologies in evaluating the squat exercise with wearable imus," *Methods of Information in Medicine*, vol. 56, no. 05, pp. 361–369, 2017.

[10] 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," *The Journal of Strength & Conditioning Research*, vol. 31, no. 8, pp. 2303–2312, 2017.

[11] M. A. O'Reilly, D. F. Whelan, T. E. Ward, E. Delahunt, and B. Caulfield, "Classification of lunge biomechanics with multiple and individual inertial measurement units," *Sports Biomechanics*, vol. 16, no. 3, pp. 342–360, 2017.

[12] P. F. Dajime, H. Smith, and Y. Zhang, "Automated classification of movement quality using the microsoft kinect v2 sensor," *Computers in Biology and Medicine*, vol. 125, p. 104021, 2020.

[13] D. Whelan, E. Delahunt, M. O'Reilly, B. Hernandez, and B. Caulfield, "Determining interrater and intrarater levels of agreement in students and clinicians when visually evaluating movement proficiency during screening assessments," *Physical Therapy*, vol. 99, no. 4, pp. 478–486, 2019.

[14] M. Paulich, M. Schepers, N. Rudigkeit, and G. Bellusci, "Xsens mtw awinda: Miniature wireless inertial-magnetic motion tracker for highly accurate 3d kinematic applications," *Xsens: Enschede, The Netherlands*, pp. 1–9, 2018.

[15] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2334–2343.

[16] 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," *The Journal of Strength & Conditioning Research*, vol. 31, no. 8, pp. 2303–2312, 2017.

[17] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *Gigascience*, vol. 6, no. 5, p. gix019, 2017.