Seven Things to Know about Exercise Classification with Inertial Sensing Wearables

Vu Phan, Ke Song, Rodrigo Scattone Silva, Karin G. Silbernagel, Josh R. Baxter, and Eni Halilaj*

***Abstract*—*Objective:* Exercise monitoring with low-cost wearables could improve the efficacy of remote physical-therapy prescriptions by tracking compliance and informing the delivery of tailored feedback. While a multitude of commercial wearables can detect activities of daily life, such as walking and running, they cannot accurately detect physical-therapy exercises. The goal of this study was to build open-source classifiers for remote physical therapy monitoring and provide insight on how data collection choices may impact classifier performance. *Methods:* We trained and evaluated multi-class classifiers using data from 19 healthy adults who performed 37 exercises while wearing 10 inertial measurement units on the wrist, pelvis, thighs, shanks, and feet. We investigated the effect of sensor density, location, type, sampling frequency, output granularity, feature engineering, and training-data size on exercise-classification performance. *Results:* Exercise groups (n=10) could be classified with 96% accuracy using a set of 10 inertial measurement units (IMUs) and with 89% accuracy using a single pelvis-worn IMU. Multiple sensor modalities (i.e., accelerometers and gyroscopes), high sampling frequencies, and more data from the same population did not improve model performance, but in the future data from diverse populations and better feature engineering could. *Conclusions:* Given the growing demand for exercise monitoring systems, our sensitivity analyses, along with open-source tools and data, should reduce barriers for product developers, who are balancing accuracy with product formfactor, and increase transparency and trust in clinicians and patients. The open-source data and code are available at https://simtk.org/projects/imu-exercise.**

***Index Terms*— Wearable sensors, inertial measurement units (IMUs), deep learning, exercise classification, remote rehabilitation.**

I. Introduction

LIGHTWEIGHT wearable sensors and the quantified-self movement they awakened in the past decade continue to inspire reflection on how to turn self-tracking from a cultural phenomenon into a healthcare revolution. Similar to glucose, temperature, heart rate, and electrical activity monitors being adopted in clinical care, remote tracking of rehabilitation exercises with motion sensors should also be feasible [1-3]. By tracking adherence, inertial sensors could inform the delivery of precision rehabilitation and lower overall healthcare costs. Yet, while the formfactors characterizing commercial products continue to improve, the modeling required to turn the data into meaningful feedback remains opaque, limiting the practical use of motion sensors. Physical activity monitors, including those embedded in consumer-grade wearables, are generally accurate but limited to recognition of basic activities of daily living, such as walking and running [4, 5].

While remote rehabilitation would benefit from accurate identification of activities that have clinical implications, the feasibility of doing so using wearable sensors remains unclear. Prior classifiers have reported high accuracy based on a small number of exercises, such as lunges, squats, and heel raises [6-9]. The training data have included a small number of both upper- and lower-extremity exercises that are highly distinctive [6, 8-14], which facilitates classification. Grouping upper- and lower-extremity exercises into a single classifier may be helpful for general fitness tracking but falls short of informing the delivery of tailored physical therapy. Clinicians often prescribe a multitude of exercises targeting specific lower or upper extremity conditions. The large number of exercises and narrower movement variability pose challenges for classifier accuracy, but provide the opportunity to cluster similar exercises that we expect will increase classifier accuracy while maximizing clinical impact.

As wearable sensor systems are developed to address this clinical need, they must not only be accurate in exercise detection but also be practical to use in patient populations. For example, sensors that require frequent patient interactions to log data or charge the battery of multiple sensors increase patient burden and decrease the large-scale implementability of wearable sensors in healthcare. However, little is known about the influence of sensor density, location, type, and optimal frequency to help balance classifier accuracy with device usability (e.g., battery life, user experience). While each study uses different sensor densities, locations, and exercises, it is difficult to understand if a wrist-worn sensor is as good as a

Manuscript received on XXX. This study was funded by awards from the National Science Foundation (CBET 2145473) and National Institutes of Health (R01AR078898, R01AR072034, P50AR080581, and NICHD R37HD037985).

*Corresponding author: Eni Halilaj, PhD (ehalilaj@andrew.cmu.edu).

Vu Phan (email: vuphan@andrew.cmu.edu) and Eni Halilaj (email: ehalilaj@andrew.cmu.edu) are with the Mechanical Engineering Department, Carnegie Mellon University Pittsburgh, PA, USA.

Ke Song (email: Ke.Song@pennmedicine.upenn.edu) and Josh R. Baxter (email: josh.baxter@pennmedicine.upenn.edu) are with the Orthopaedic Surgery Department, University of Pennsylvania, Philadelphia, PA, USA.

Rodrigo Scattone Silva (email: r.scattone@outlook.com) is with the Postgraduate Program in Rehabilitation Sciences, Federal University of Rio Grande do Norte, Santa Cruz, Brazil.

Karin G. Silbernagel (email: kgs@udel.edu) is with the Physical Therapy Department, University of Delaware, Newark, DE, USA.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3368042

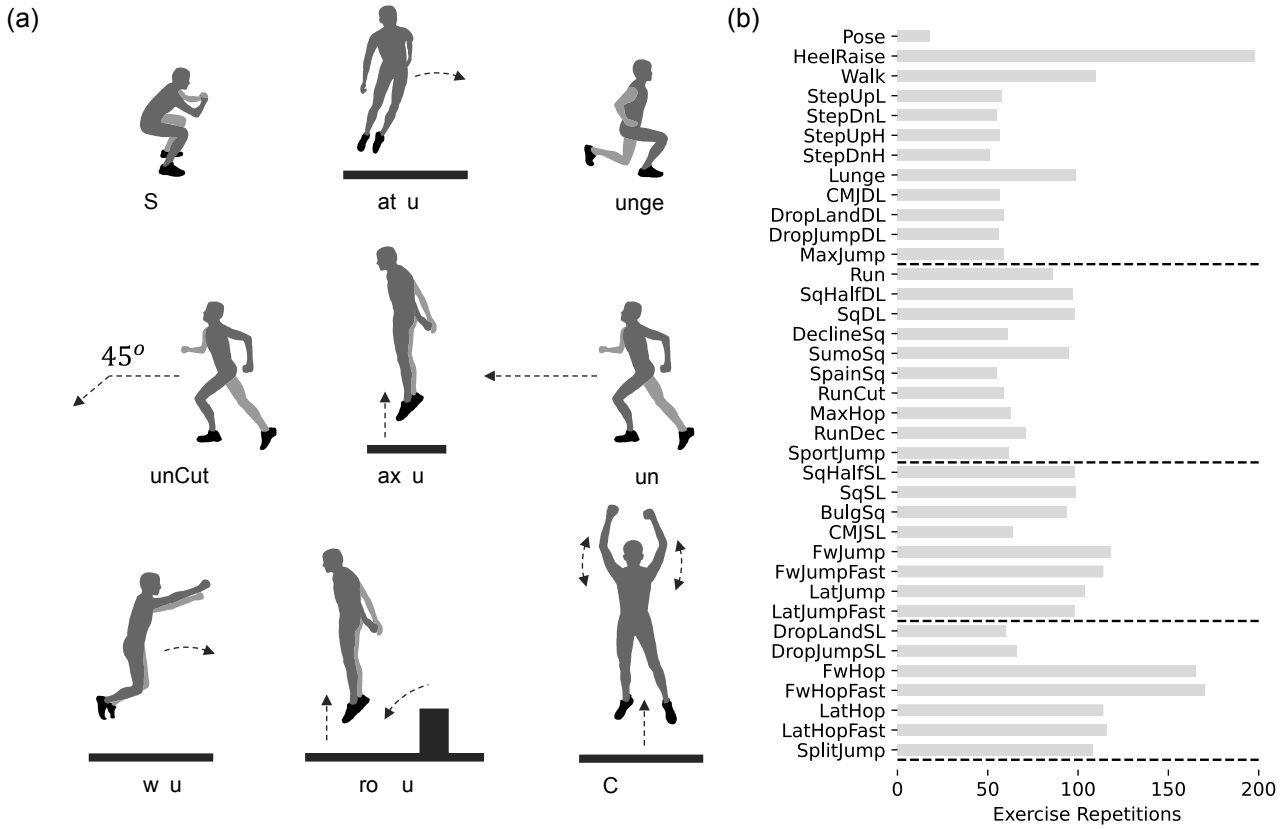

Fig. 1. Data Collection Overview. (a) Examples of a subject performing a few of the lower-extremity exercises captured in this study. (b) The distribution of exercise repetitions across the entire dataset of 19 subjects. Exercises were divided into 4 modules based on their intensity, with breaks of 3 – 5 minutes between each module. Data were collected in the order presented here. Definitions of the exercise acronyms can be found in Table I.

pelvis-worn one, or if a pelvis-worn one is as good as multiple segment-specific sensors. Recent work considered single-sensor placements at the wrist [6, 17], forearm [14, 15], upper arm [18], lower back [19, 20], and shank [21], but it is difficult to draw conclusions on which sensor location is optimal due to differences in the exercises these studies included. Further, as most studies focused on performance metrics, few translatable insights have been provided to help researchers, clinicians, and product developers choose parameters that are most practical when implementing a monitoring system. For example, it is unclear if power-hungry gyroscopes are essential for classifier performance over accelerometry alone. It is also unknown if sampling frequency can be lowered to extend battery life without sacrificing performance. Another question is if the models could be enhanced with more data and better feature engineering, or if they have reached theoretical limits of performance. One large focus in movement biomechanics is kinematics estimation with inertial measurement units (IMUs) [22, 23]. However, exercise classification models have in the past been trained using accelerometry and gyroscope data extracted directly from the sensors, instead of estimated kinematics, which can be interpreted as domain-informed feature engineering. In many applications, feature engineering is a laborious step that can ultimately improve classifier accuracy, but it remains unclear if this applies to exercise classification.

The primary goal of this work was to build open-source exercise classification models and perform sensitivity analyses that would inform the development of remote physical therapy monitoring systems and data collection protocols, democratizing the process for all researchers, clinicians, and product-development entities. We first report classifier performance across 37 lower-extremity load-bearing exercises and use a data-driven approach to cluster similar exercises. We then systematically investigate the impact of sensor density, location, type, and sampling rate. Last, we investigate if more data and better feature engineering would move these classifiers toward theoretical limits of performance. In addition to the code, we make all the data publicly available.

II. Methods

Nineteen healthy subjects (9 males:10 females; age: 25 ± 5 years; body-mass index: 24.1 ± 2.4 kg/m2) were recruited to participate in this study after we obtained approval from the Institutional Review Board at the University of Pennsylvania and informed consents [24]. None of the participants had self-reported injuries in the past 6 months. They were instructed to perform 37 lower-extremity exercises while following visual demonstrations by a physical therapist. At least three successful repetitions were recorded for each exercise. Based on the intensity of exercises, we divided data collection into 4 modules, each containing 7 to 10 exercises (Fig. 1 and Table I).

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3368042

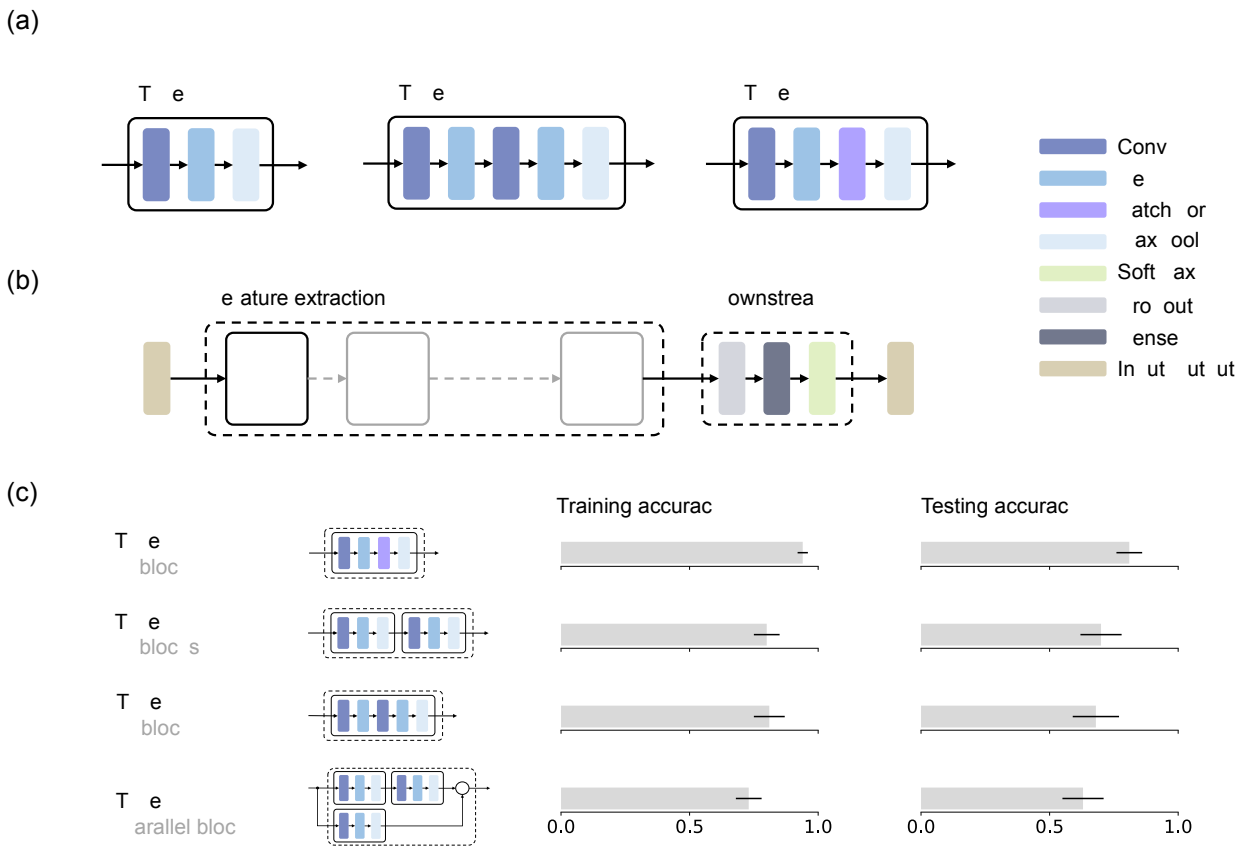

Fig. 2. Model Architecture Selection. (a) Three basic combinations of common CNN components were used: Type 1 has a Conv1D layer followed by MaxPool1D layer, Type 2 has two Conv1D layers and a MaxPool1D layer, and Type 3 has a Conv1D layer followed by a BatchNorm and a MaxPool1D layers. These were used as building blocks across different models. (b) The sequential architecture of the deep learning model with various building blocks for feature extraction. The blocks were added to the feature extractor until no further improvement was observed. While the structure and hyperparameters of the feature extraction block were fine-tuned, those of the downstream block remained the same. (c) The deep learning model with a Type-3 block outperformed the others. An additional model that used a parallel architecture of Type-1 blocks was applied. However, the performance of this model was relatively poor compared to its sequential counterparts.

Subjects took breaks of 3 – 5 minutes after each module to prevent any possible confounding effects of physical exhaustion.

We tracked movement with both inertial sensors (Opal, APDM Wearable Technologies, USA) and marker-based motion capture (Raptor Series, Motion Analysis Corporation, USA), with the latter used to obtain ground-truth movement. Ten IMUs were attached to the subject's chest, pelvis, wrists, thighs, shanks, and feet (Fig. 4) to measure three-dimensional linear accelerations via accelerometry and angular velocities via gyroscopes. These data were used as input to train the deep learning models, with the labels indicating exercise type recorded by the research team. The 12-camera marker-based motion capture system recorded the motion of 31 retro-reflective markers placed on the lower body. The purpose of using marker-based motion capture was to obtain idealized joint kinematics and determine whether they can help improve classifier performance. IMU and motion capture data were sampled at 100 Hz. Three in-floor force platforms (BP600900, AMTI Force and Motion, USA) concurrently recorded ground reaction forces at 1000 Hz. Data from different modalities were synchronized by a syncing device (V2 Sync Box, APDM Wearable Technologies, USA). Marker-based motion capture data were filtered with a low-pass 4th order Butterworth filter with a cut-off frequency of 6 Hz before being used to perform inverse kinematics. To segment the linear acceleration, angular velocity, and marker-based motion capture data into individual exercise repetitions, we used the force plate data. Exercise repetitions were then normalized and resampled (0 – 100%).

We first built convolutional neural network (CNN) models with different architectures to classify 37 exercises, before performing a cluster analysis to determine if the exercises could be further simplified into fewer groups. Building blocks consisted of rectified linear activated convolutional layers, batch normalization layers, and maximum pooling layers (Fig. 2). The number of building blocks (1 – 3) was a hyperparameter, along with batch size (16 – 128), convolutional outputs (32 – 256), and pool size (2 – 8), which were fine-tuned though a grid search. The convolutional kernel size and stride grid did not affect performance and were kept constant at 4 and 1, respectively. We used leave-one-subject-out cross validation (LOSOCV) for both hyperparameter tuning and performance evaluation, splitting the data into train, validation, and test sets.

We carried out a cluster analysis using data from all ten IMUs to determine which exercises could be grouped together for lighter-weight models. Since exercise with similar

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3368042

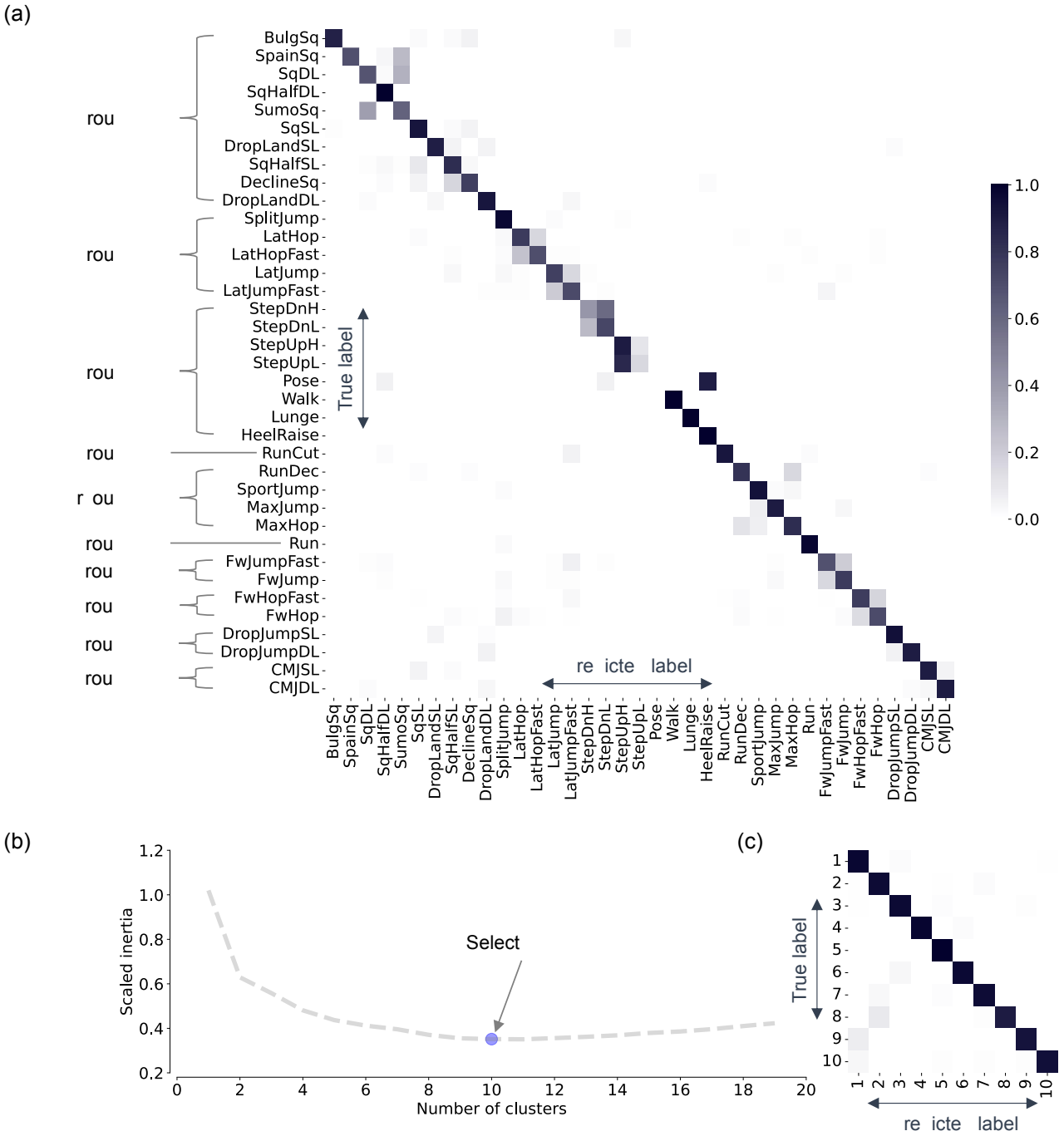

Fig. 3. Full-Sensor Model Performance. (a) The deep learning model reached an accuracy of 81% in identifying 37 specific exercises. The confusion matrix illustrates that misclassifications were due to exercises with similar movements, such as forward jump versus forward jump fast or forward hop versus forward hop fast. (b) The elbow method suggested that the optimal number of clusters for k-mean clustering was 10. (c) The cluster-specific classifier was more accurate than the exercise-specific classifier, with accuracy increasing from 81% to around 96%. Definitions of the exercise acronyms can be found in Table I.

movements could result in similar clinical outcomes (e.g., joint load), this grouping of exercises practically helps reduce patient burden, while optimizing classifier performance. K-means was used to cluster exercises into groups. We used the elbow method on a curve of scaled inertia to determine the optimal number of clusters, with the k ranging from 1 to 19. A sensitivity analysis of the initialization parameters revealed they had little bearing on the final results. Inertial data from all the repetitions of the same exercise were averaged to obtain 37 samples representing 37 exercises. Each sample was normalized and resampled as noted earlier before being input into the cluster analysis. Once the exercise groups were determined, we re-trained and re-evaluated the models for classification of these groups. Here we report model performance in terms of accuracy, precision, recall, and F1 score. For generalizability and rigor in model evaluation, we

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3368042

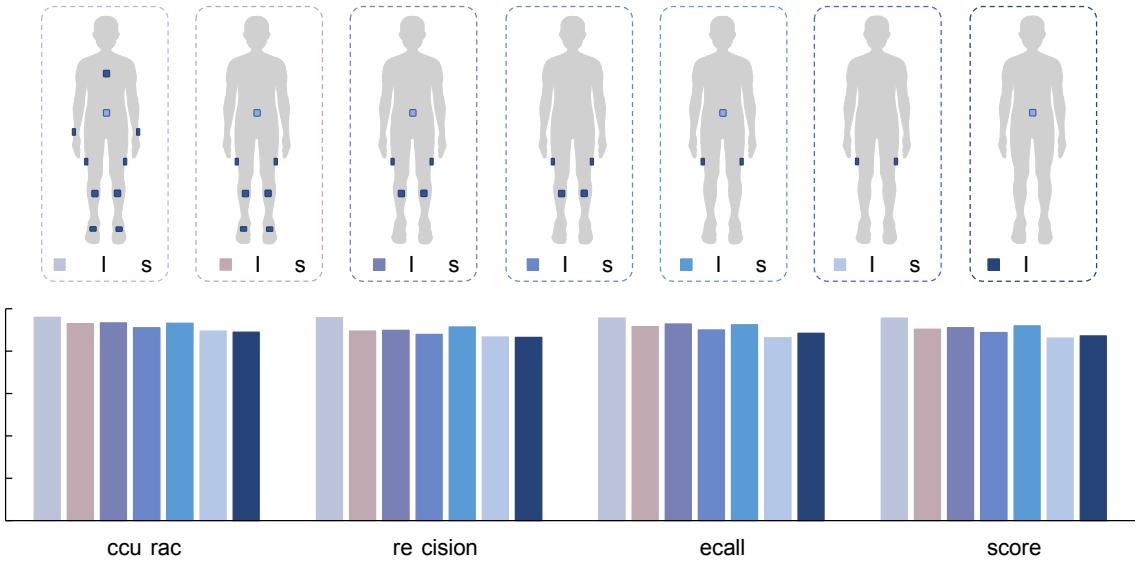

Fig. 4. Model Performance by Sensor Density: Exercise Groups. The model using data from 3 lower-body sensors placed on pelvis, and thighs achieved a performance that was similar to that of the full sensor set. With fewer than 3 sensors, performance dropped progressively. Of the single-sensor configurations, the pelvis one achieved the highest performance. Numerical results can be found in Table S1 (supplementary document).

ensured that all the data from a subject were included solely in the training, validation, or test set.

In addition to training models with the full set of data from 10 IMUs, we also trained sparser models with data from 7, 5, 4, 3, 2, and 1 IMUs. For the single-sensor models, the IMUs placed on right wrist, chest, pelvis, and right thigh were tested due to their convenience for consumer products (Fig. 5a and 9a). For double-sensor settings, the sensors attached to chest and right thigh, both thighs, right wrist and right thigh, and right thigh and right shank (Fig. 5b and 9b) were tested.

To address questions around battery life, we compared models built with different sensor types and data sampling frequencies. Specifically, for the full-, single-, and double-sensor configurations, the models were trained and tested with data from the accelerometer only and gyroscope only, in addition to the original model that included both. To test model sensitivity to sampling frequency, the IMU data were downsampled to 50, 25, 20, 10, and 5 Hz and full-sensor models were retrained.

Last, we explored whether these models could be further improved via feature engineering and additional training data. To show that domain knowledge can boost model performance, we obtained kinematics from the marker-based motion capture data and trained models using only kinematics as the input feature. Marker-based kinematics represent idealized kinematics, which could theoretically be derived from IMU data via feature engineering. To gauge how model performance changes with the number of subjects included in the training dataset, learning curves were generated using 2 to 19 subjects. An inverse power law curve fitting approach [32, 33] was applied to predict how performance would increase with hypothetical sample sizes and what the theoretical limits of these models are.

All data and codes of this work will be available on SimTK (link: https://simtk.org/projects/imu-exercise).

III. Results

A classifier built to predict exercise groups (i.e., clusters) was more accurate than one built to predict individual exercises. The optimal number of exercise clusters was 10 (Fig. 3b), and similar exercises were grouped together (Fig. 3a). For example, all the squats were grouped into the first group. Using linear acceleration and angular velocity data from 10 IMUs, exercise group could be predicted with nearly 96% accuracy (Fig. 3c and 4), while the 37 individual exercise with 81% accuracy (Fig. 2c, 3a, and 8). Most exercises misclassified by the individual exercise model had similar movements and were typically clustered under the same group. Examples include forward jump versus forward jump fast and lateral hop versus lateral hop fast (Fig. 3a).

Classification of exercise clusters was less sensitive to sensor number and location than classification of individual exercises. Reducing the sensor set reduced cluster-specific model performance only minimally, from 96% to 93%, 91%, 93%, 89% for the 5-, 4-, 3-, 2-IMU systems, respectively (Fig. 4). Single-IMU models achieved accuracies of 89% for the pelvis and 75% for the wrist (Fig. 5). Sensor density and location, however, had a major impact on classification of individual exercises, with accuracy dropping by more than 20%, from 81% to 61%, when a single pelvis-worn IMU was used compared to 10 IMUs (Fig. 8 and 9).

Accelerometry-based models were as accurate as those using both accelerometer and gyroscope data, and sampling frequency could be reduced to 20 Hz without affecting model accuracy (Fig. 6). With the full 10-IMU sensor set, a model using accelerometer data achieved similar performance to that using both accelerometer and gyroscope data sources (Fig. 6a). In addition, models trained with data that were downsampled to 50 or 25 Hz performed as well as the model trained with the

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3368042

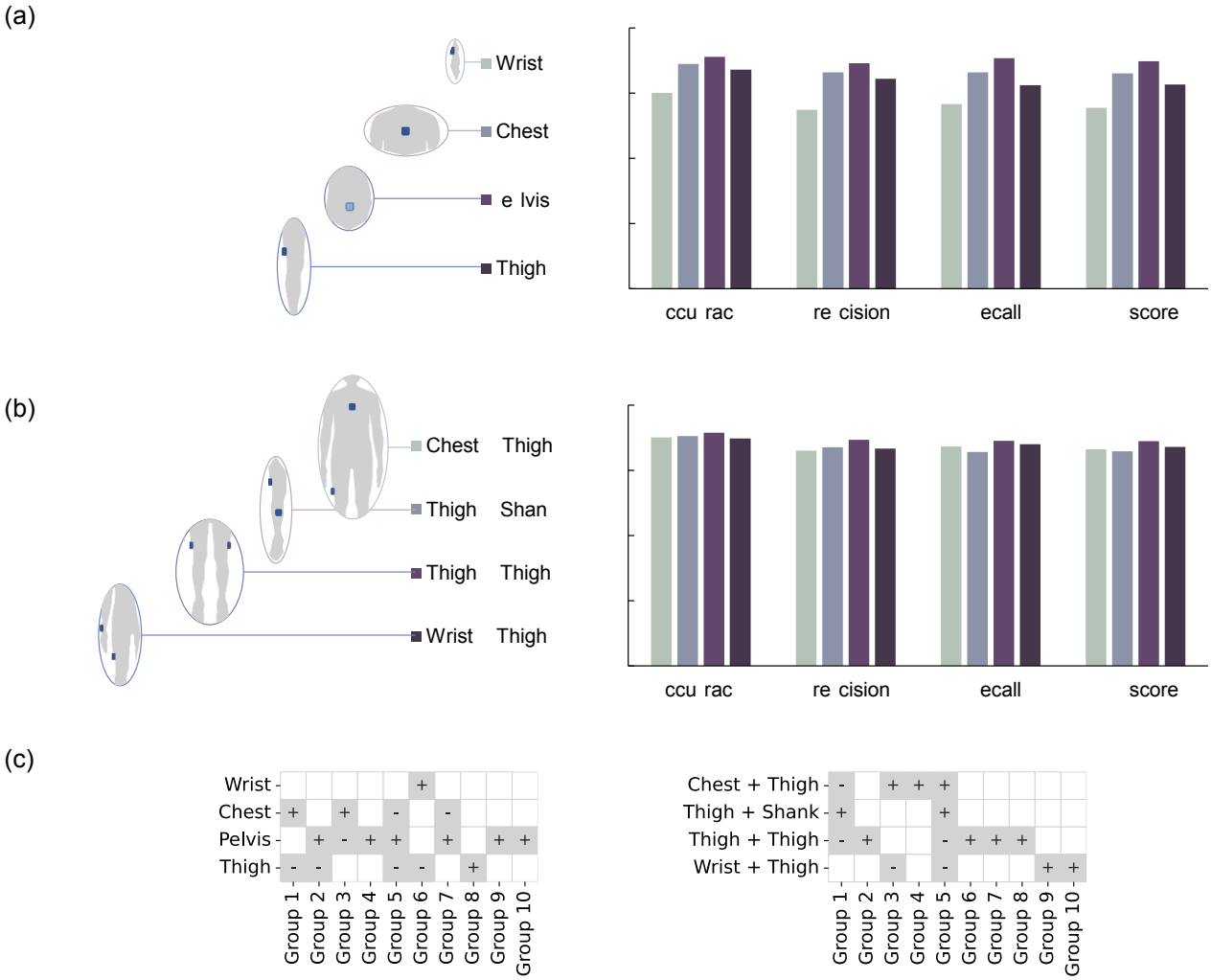

Fig. 5. Model Performance by Sensor Location: Exercise Groups. (a) Among single-sensor configurations, a sensor placed on the pelvis achieved the highest performance overall (i.e., 89% in accuracy and 0.87 in F1-score). (b) A combination of sensors placed on both thighs achieved the highest performance across two-sensor configurations (i.e., 89% in accuracy and 0.86 in F1-score). (c) The optimal sensor placements (for either single- or double-sensor settings) varied by specific exercise, as illustrate in the botto anels, where " " re resents the best erfor ance er exercise, an "-" in icates less than or e ual to % ecreases in erfor an ce co a re to the best erfor an ce. Numerical results of (a) and (b) can be found in Tables S2 and S3 (supplementary document), respectively.

original data recorded at 100 Hz. Further reducing the sampling frequency to 20 Hz reduced performance by less than 3% (Fig. 6b). These trends were similar for the exercise-specific models (Fig. 10).

Classification of exercise clusters was also less sensitive to training sample size and feature engineering than classification of individual exercises (Fig. 7). Our sample size analysis indicated that adding 132 subjects would increase the accuracy of identifying exercise clusters from 96% to 99%. Kinematics data increased accuracy from 96% to 98% (Fig. 7a). On the other hand, more subjects (n = 2322) would improve the accuracy of exercise-specific models from 81% to 95%, when using accelerometry and gyroscope data (Fig. 7b). Kinematics would help improve the performance from 81% to 88% when data from only 19 subjects are available. To reach 95% accuracy, a kinematics-based model would require only 237 subjects. Generally, the accuracy of these models is theoretically bound at 98%.

IV. Discussion

The goal of this study was to develop open-source models for exercise classification and carry out detailed sensitivity analyses on sensor density, location, type, sampling frequency, feature engineering, and training data size. We found that deep learning classifiers could predict 10 exercise groups with an accuracy of 96% and 37 individual exercises with an accuracy of 81%. Ten full-body sensors and three placed on the lower body led to comparable performance (i.e., less than 3% difference). Of single-sensor systems, one worn on the pelvis led to the best performance. Accelerometry data alone performed as well as gyroscope and accelerometry together, and when these modalities are jointly used to derive kinematics, model performance improved only marginally. Our sample size analysis indicated that more data from the same population improves classification of individual exercises, but not exercise groups.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3368042

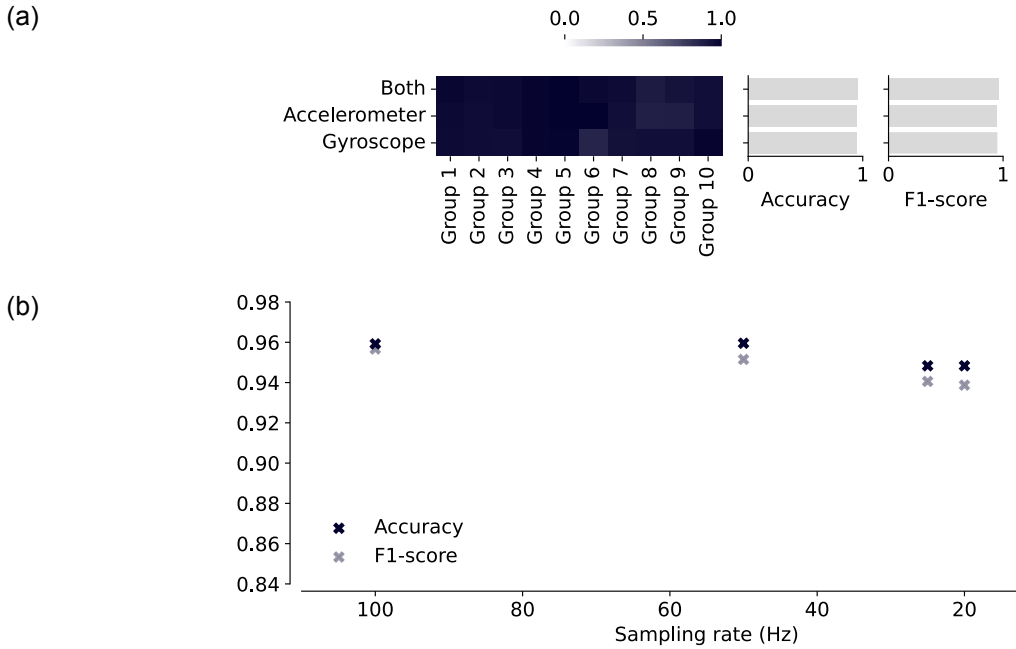

Fig. 6. Model Performance by Sensor Type and Sampling Frequency: Exercise Groups. (a) The model trained with accelerometer data provided equivalent performance as that trained with gyroscope data or both. (b) Overall, reducing the rate at which the sensor data are sampled by the factor of up to 5 (i.e., 20 Hz) caused less than 3% decrease in performance. The performance decreased more noticeably as the sampling rate was reduced to 10 Hz and 5 Hz. Numerical results of (a) and (b) can be found in Tables S4 and S5 (supplementary document), respectively.

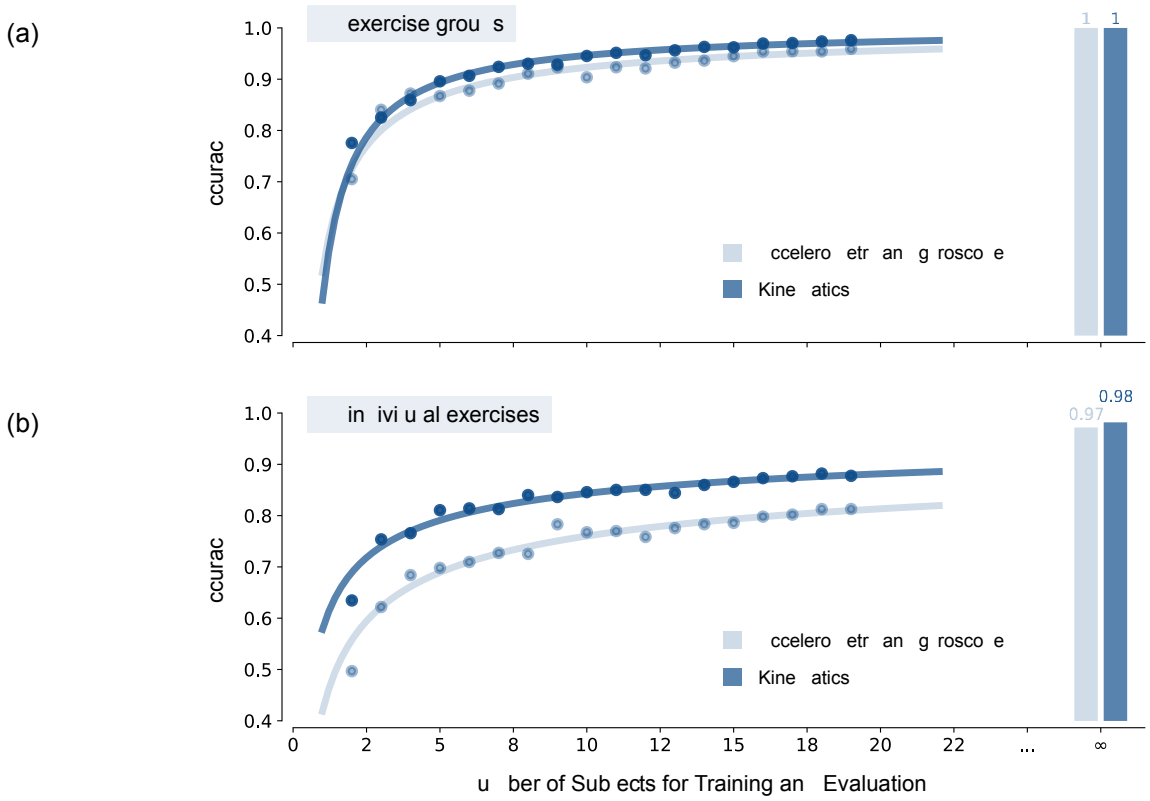

Fig. 7. Model Performance with Sample Size and Feature Engineering. A learning curve was derived after training and evaluating models with 2 to 19 subjects (i.e., dots in the figures). An inverse power law [32, 33] estimated how the models could improve with more data (i.e., solid lines), akin to sample size calculations in inferential statistics. (a) This sample size analysis predicted that group-specific models can theoretically reach accuracies of 100%. (b) Exercise-specific models can reach accuracies of 97% and 98% for the baseline and kinematics model, respectively, but data from thousands of subjects would be required. Numerical results can be found in Table S6 (supplementary document).

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3368042

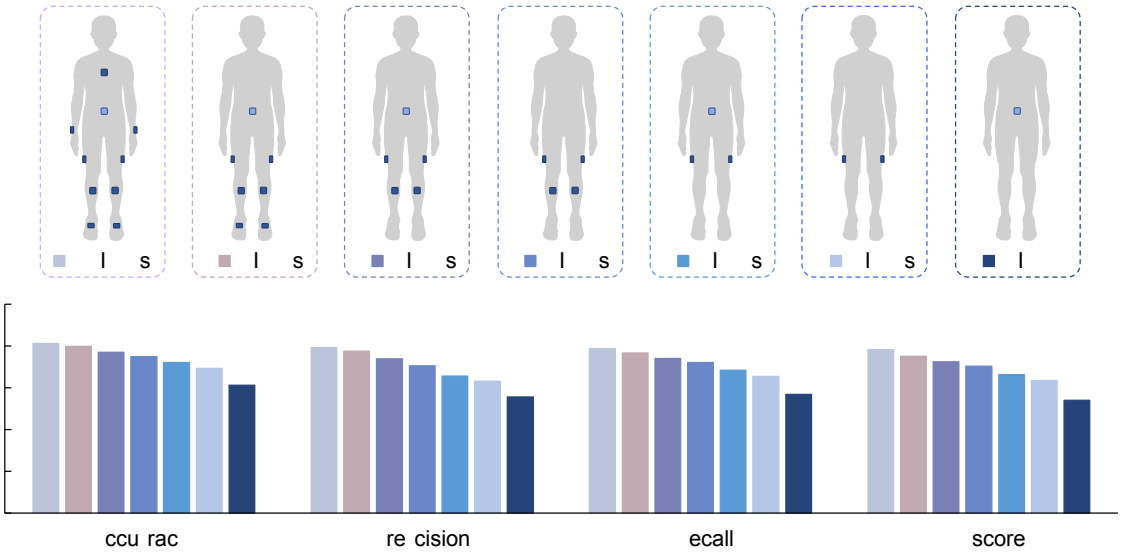

Fig. 8. Model Performance by Sensor Density: Individual Exercises. Unlike the cluster-specific classifier, whose performance maintained high accuracy even with a single sensor, the accuracy of the exercise-specific model dropped with reduced sensor density. Numerical results can be found in Table S1 (supplementary document).

Our work suggests that grouping exercises by similarity is not only practically beneficial, but also clinically relevant. The clustering approach, which increased model performance from 81% to 96%, automatically grouped exercises into 10 categories with considerable affinity in terms of how they load joints of interest. For example, a prior analysis has shown that the patellofemoral joint is loaded similarly by a two-leg countermovement jump (i.e., CMJDL) and one-leg countermovement hop (i.e., CMJSL) [24] and our clustering analysis grouped these activities together. The goal of remote exercise prescription and monitoring programs is to track how joints, ligaments, tendons, and muscles are loaded, and if a few exercises bear similar clinical significance in this context, then distinguishing among them at the expense of accuracy is not worthwhile. This grouping has the added advantage of generating equally accurate models with sparser sensor sets, making it a more viable solution for translation. A pelvis-worn sensor, which tracks the body's center of mass, stands out among single-sensor configurations with an accuracy of 86% in predicting exercise clusters—a decrease of only 9% from the model that uses the full set of 10 segment-specific sensors.

Our estimations also indicate that by cutting sampling frequency and gyroscope data, we can prolong battery life without sacrificing accuracy. The gyroscope consumes at least 6 times more energy than the accelerometer [25, 26] but does not substantially improve accuracy. Typically, gyroscope data are more informative in highly dynamic movements, which is not the case in rehabilitation exercises to avoid injuries for patients. However, the complementary nature of these two sensors can be harnessed by using domain knowledge to further improve classifier performance, as noted below. As movement frequency for physical therapy exercises is likely below 10 Hz, a sampling frequency of 20 Hz should be the lowest sampling frequency that guarantees good performance based on the Nyquist theorem [27-30]. At the cost of less than 3% reduction

TABLE I
LIST OF EXERCISES IN THE DATASET (IN THE ORDER OF DATA COLLECTION) AND THEIR ABBREVIATIONS

Module	Exercise	Abbreviation
1	Static pose	Pose
	Heel raises	HeelRaise
	Walk	Walk
	Step up (low step)	StepUpL
	Step down (low step)	StepDnL
	Step up (high step)	StepUpH
	Step down (high step)	StepDnH
	Lunges	Lunge
	2-leg counter-movement jump	CMJDL
	2-leg drop-landing	DropLandDL
	2-leg drop-and-jump	DropJumpDL
	2-leg maximal forward jump	MaxJump
2	Run	Run
	2-leg squat (half depth)	SqHalfDL
	2-leg squat (full depth)	SqDL
	Decline squat	DeclineSq
	Sumo squat	SumoSq
	Spanish squat	SpainSq
	Run-and-cut	RunCut
	1-leg maximal forward hop	MaxHop
	Run-and-stop	RunDec
	Sports movement jump	SportJump
3	1-leg squat (half depth)	SqHaflSL
	1-leg squat (full depth)	SqSL
	Bulgarian squat	BulgSq
	1-leg counter-movement hop	CMJSL
	2-leg repeated forward jumps (regular)	FwJump
	2-leg repeated forward jumps (fast)	FwJumpFast
	2-leg repeated lateral jumps (regular)	LatJump
	2-leg repeated lateral jumps (fast)	LatJumpFast
4	1-leg drop-landing	DropLandSL
	1-leg drop-and-hop	DropJumpSL
	1-leg repeated forward hops (regular)	FwHop
	1-leg repeated forward hops (fast)	FwHopFast
	1-leg repeated lateral hops (regular)	LatHop
	1-leg repeated lateral hops (fast)	LatHopFast
	Alternating split jumps	SplitJump

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3368042

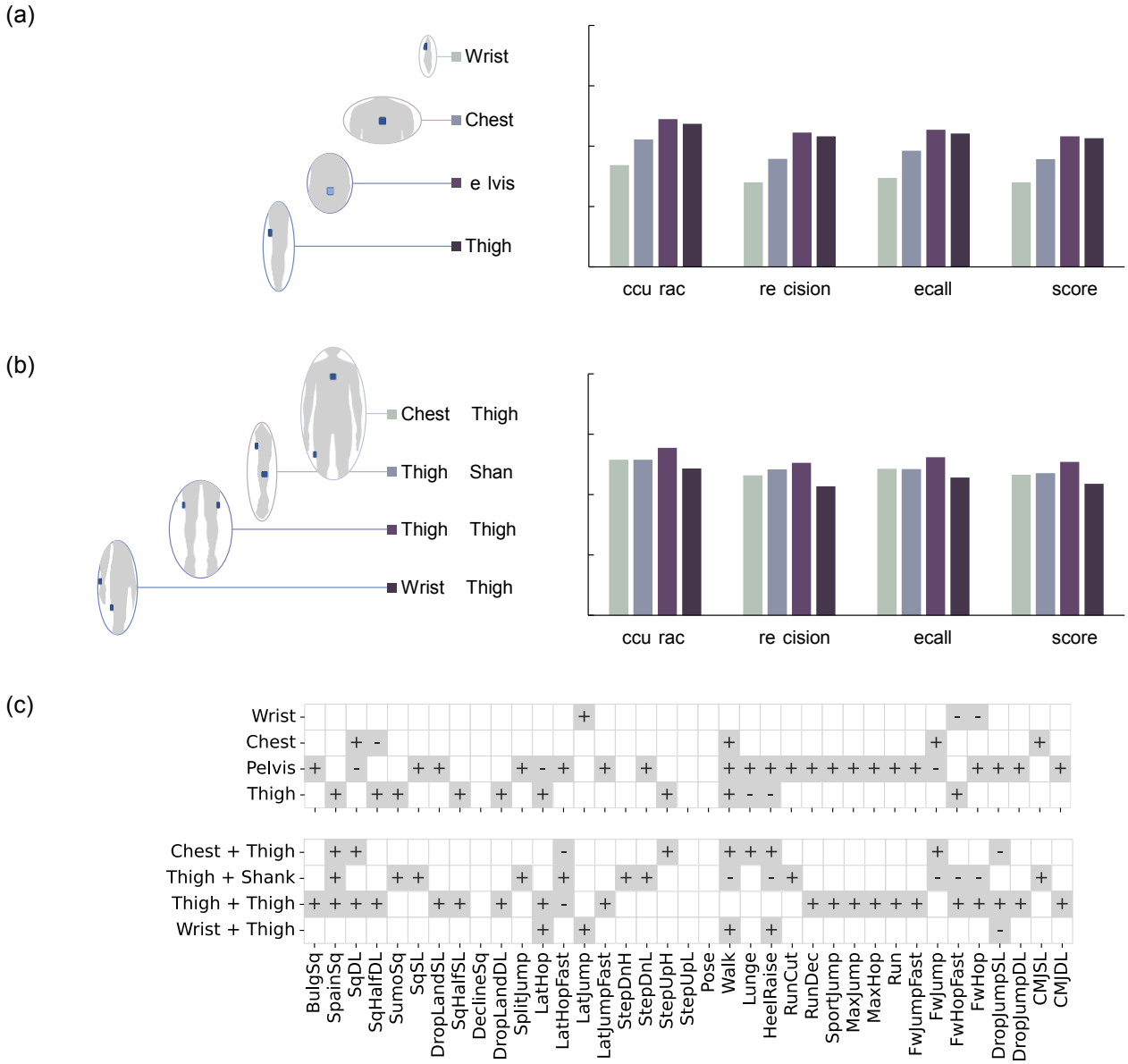

Fig. 9. Model Performance by Sensor Location: Individual Exercises. (a) Among single-sensor configurations, a sensor placed on the pelvis achieved the highest performance overall (i.e., 61% in accuracy and 0.54 in F1-score). (b) A combination of sensors placed on both thighs achieved the highest performance across two-sensor configurations (i.e., 69% in accuracy and 0.64 in F1-score). (c) The optimal sensor placement (for either single- or double-sensor settings varie b s ecific exercise, as illustrate in the botto anel, where " " re resents the best erfor an ce er exercise, an "-" in icates less than or e ual to % ecreases in erfor an ce co a re to the best erfor an ce efinitions of the exercise acronyms can be found in Table I. Numerical results of (a) and (b) can be found in Tables S2 and S3 (supplementary document), respectively.

in performance, this sampling rate theoretically improves battery life up to 5 times, which is significant for out-of-clinic monitoring applications [31].

Moving forward, we have identified two areas for further investigation to improve exercise classification models: having more diverse data and better feature engineering. While more data from the same population do not seem to be beneficial, diverse data, from older adults or those with musculoskeletal diseases, could further improve model generalizability. Good feature engineering is a more judicious and cost-efficient way to improve performance compared to obtaining more data. Models trained on kinematics, as opposed to accelerations and angular velocities measured directly by the sensors, could improve classification of individual exercises and exercise groups by small margins. Two points are important to consider in this context. First, to build the kinematics-based classifier, instead of estimating kinematics from IMU data, we used ground-truth kinematics from the marker-based motion capture system. This model represents a best-case scenario for how accurately kinematics can be estimated in natural environments. In reality, estimation of kinematics with IMUs is an ongoing technical challenge, albeit one that is receiving considerable attention. Second, it is important to note that a heuristics-guided approach could be better than a purely data-driven one at training deep learning models when the kinematics data are available. For example, an exercise can be detected based on

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3368042

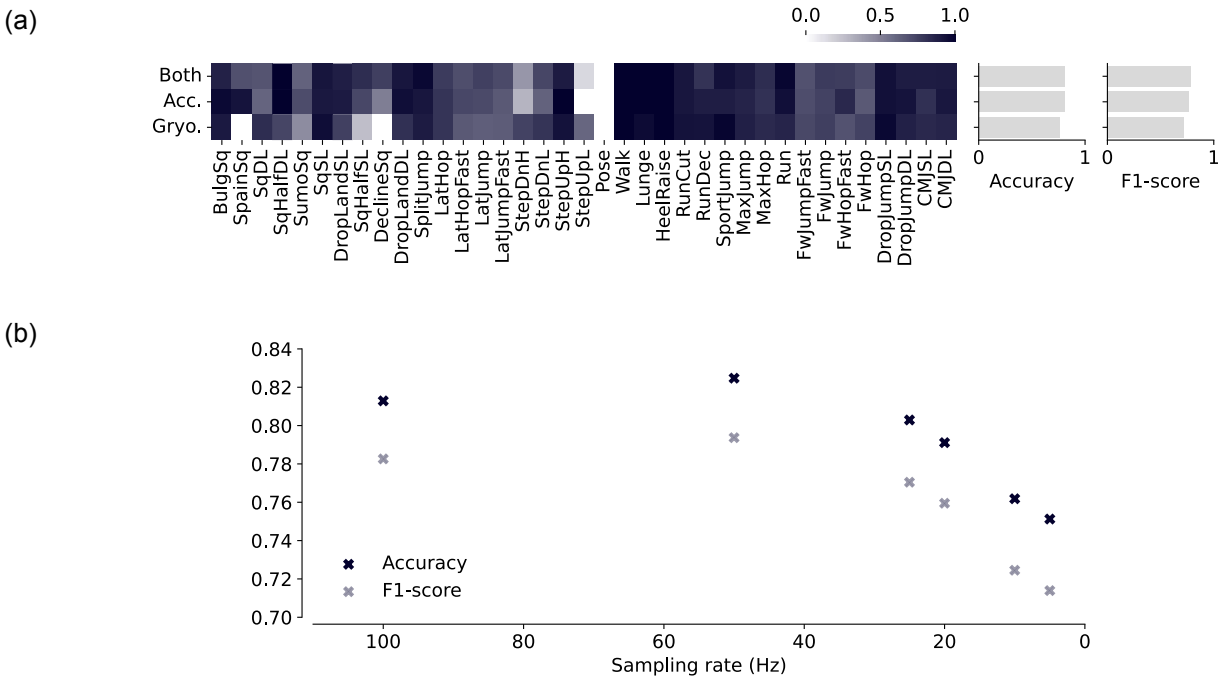

Fig. 10. Model Performance by Sensor Type and Sampling Frequency: Individual Exercises. (a) The accelerometer data was more informative in detecting exercises than the gyroscope data. (a) Overall, reducing the rate at which the sensor data are sampled by the factor of up to 5 (i.e., 20 Hz) almost did not alter the performance (i.e., less than 3% in difference). The performance decreased noticeably as the sampling rate was reduced to lower than 20 Hz. Numerical results of (a) and (b) can be found on Tables S4 and S5 (supplementary document), respectively.

joint- or segment-specific heuristics. While such an approach is not feasible with accelerometry and angular velocity data, it is the most intuitive and likely the most accurate when kinematics data are available. The improvement in accuracy we report with a deep learning approach therefore represents a worst-case scenario. These factors that contribute to classification performance should be carefully considered in the broader context of implementing wearable sensors in clinical populations. For example, using a kinematics-based classifier is likely to marginally improve performance but at the expense of additional sensors and data streams that increase the patient burden beyond the performance benefit. This cost-benefit analysis should be considered for each application and likely will differ based on the patient cohort.

A few limitations are important to consider when interpreting these results. First, we used only data from healthy adults to train the models. However, we expect differences across individuals to be smaller than differences across exercises, which makes the classifier highly likely to generalize to patient populations, although this remains to be demonstrated. Second, the data were collected in a controlled laboratory environment and may be cleaner than natural-environment data. Future efforts may address this potential limitation through the use of transfer learning, where a small set of natural-environment data can be used to fine-tune the models. Third, we did not emphasize the practical challenge of detecting non-exercise movements as we envisioned the use of a wearable system only during exercising sessions. Last, exercise segmentation was carried out by experts here since creating an end-to-end system is an engineering task that was beyond the scientific scope of this work. Prior work focusing on automated segmentation from inertial data has achieved accuracies of 96% [34, 35].

V. Conclusion

We summarize our findings in the following seven lessons and hope that would help fuel progress and more informed translation.

1) We should aim for clinical relevance over inconsequential granularity. Understanding how much time each patient spends in each of the 10 exercise clusters gives clinicians sufficient insight on joint loading.
2) Exercise group classification can be performed with a single sensor, while detection of individual exercises needs up to five sensors on the lower body.
3) A pelvis, chest, and thigh-worn sensors are the most informative for lower-extremity exercise monitoring, while a wrist-worn sensor is the worst.
4) The gyroscope can be removed without sacrificing performance. We recommend leaving out the power-hungry gyroscope.
5) Sampling frequency can be reduced to 20 Hz to prolong battery life and preserve sensor memory that will extend monitoring capabilities and decrease patient burdens in out-of-clinic environments.
6) More data from the same population may not be beneficial. Instead, we suggest diversifying the training data to include a diverse sense of patients and possibly employ transfer learning.

7) Better feature engineering could improve classifier accuracy, but that improvement should be weighted against the added burden to patients resulting from multi-sensor system requirements.

The open-source code should help verify our conclusions on additional datasets, increase product transparency, and build trust in patients. Our open-source provides normative wearables data across a wide range of lower-extremity rehabilitation exercises that are used to treat and screen individuals recovering from musculoskeletal injuries, including anterior cruciate ligament, patellar tendon, and Achilles tendon pathology.

APPENDIX

A. Lower-extremity Exercises

Exercises included in the dataset and their abbreviations can be found in Table I.

B. Performance of Exercise-Specific Models

In addition to outcomes of the group-specific models reported in the main content, here, we present performance of models when classifying 37 individual exercises, see Figs. 8, 9, and 10.

ACKNOWLEDGMENT

The authors would like to thank Todd Hullfish, Audrey Lehneis, and Liliann Zou at University of Pennsylvania for their assistance with data collection and processing; Neel Joshi at Carnegie Mellon University for helping with data processing; and the reviewers for constructive suggestions to improve this manuscript.

REFERENCES

[1] Bayoumy, K., *et al.*, "Smart wearable devices in cardiovascular care: where we are and how to move forward," *Nature Reviews Cardiology*, 1(8), 581-599, 2021.

[2] Guk, K., *et al.*, "Evolution of wearable devices with real-time disease monitoring for personalized healthcare," *Nanomaterials*, 9(6), 813, 2019.

[3] Prieto-Avalos, G, *et al.*, "Wearable devices for physical monitoring of heart: a review," *Biosensors*, 12(5), 292, 2022.

[4] Fuller, D., *et al.*, "Predicting lying, sitting, walking and running using Apple Watch and Fitbit data," *BMJ Open Sport & Exercise Medicine*, 7(1), e001004, 2021.

[5] Lara, O. D. and Labrador M. A., "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, 15(3), 1192-1209, 2012.

[6] Conger, S. A., *et al.*, "Objective Assessment of Strength Training Exercises using a Wrist-Worn Accelerometer," *Medicine and science in sports and exercise,* 48(9), 1847-1855, 2016.

[7] O'Reilly, M. A., *et al.*, "Technology in strength and conditioning tracking lower-limb exercises with wearable sensors," *The Journal of Strength & Conditioning Research*, 31(6), 1726-1736, 2017.

[8] Morris, D., *et al.*, "RecoFit: using a wearable sensor to find, recognize, and count repetitive exercises," *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3225-3234, 2014.

[9] O'Reilly, M., *et al.*, "Wearable inertial sensor systems for lower limb exercise detection and evaluation: a systematic review," *Sports Medicine*, 48, 1221-1246, 2018.

[10] Alfakir, A., *et al.*, "Detection of Low Back Physiotherapy Exercises with Inertial Sensors and Machine Learning: Algorithm Development and Validation," *JMIR Rehabilitation and Assistive Technologies*, 9(3), e38689, 2022.

[11] García-de-Villa, S., *et al.*, "Simultaneous exercise recognition and evaluation in prescribed routines: Approach to virtual coaches," *Expert Systems with Applications,* 199, 116990, 2022.

[12] Crema, C., *et al.*, "Characterization of a wearable system for automatic supervision of fitness exercises," *Measurement*, 147, 106810, 2019.

[13] Tian, J., *et al.*, "Wearable IMU-based gym exercise recognition using data fusion methods," *In the Fifth International Conference on Biological Information and Biomedical Engineering*, pp. 1-7, 2021.

[14] Um, T. T., *et al.*, "Exercise motion classification from large-scale wearable sensor data using convolutional neural networks," *In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2385-2390, 2017.

[15] O'Reilly, M. A., *et al.*, "Classification of deadlift biomechanics with wearable inertial measurement units," *Journal of biomechanics*, 58, 155-161, 2017.

[16] Allahbakhshi, H., *et al.*, "The key factors in physical activity type detection using real-life data: a systematic review," *Frontiers in physiology*, 10, 75, 2019.

[17] Crema, C., *et al.*, "IMU-based solution for automatic detection and classification of exercises in the fitness scenario," *In 2017 IEEE Sensors Applications Symposium (SAS)*, pp. 1-6, 2017.

[18] Bavan, L., *et al.*, "Adherence monitoring of rehabilitation exercise with inertial sensors: A clinical validation study," *Gait & posture*, 70, 211-217, 2019.

[19] O'Reilly, M., *et al.*, "Evaluating squat performance with a single inertial measurement unit," *In 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 1-6, 2015.

[20] Whelan, D., *et al.*, "Evaluating performance of the single leg squat exercise with a single inertial measurement unit," *In Proceedings of the 3rd 2015 Workshop on ICTs for improving Patients Rehabilitation Research Techniques*, pp. 144-147, 2015.

[21] Argent, R., *et al.*, "The importance of real-world validation of machine learning systems in wearable exercise biofeedback platforms: A case study," *Sensors*, 21(7), 2346, 2021.

[22] Bonnet, V., *et al.*, "Monitoring of hip and knee joint angles using a single inertial measurement unit during lower limb rehabilitation," *IEEE Sensors journal*, 16(6), 1557-1564, 2015.

[23] O'Donovan, K. J., *et al.*, "An inertial and magnetic sensor-based technique for joint angle measurement," *Journal of biomechanics*, 40(12), 2604-2611, 2007.

[24] Song K., *et al.*, "Patellofemoral joint loading progression across 35 weight-bearing rehabilitation exercises and activities of daily living," *American Journal of Sports Medicine*, in press, 2023.

[25] Liu, Q., *et al.*, "Gazelle: Energy-efficient wearable analysis for running," *IEEE Transactions on Mobile Computing*, 16(9), 2531-2544, 2016.

[26] Zhang, L., *et al.*, "Accelword: Energy efficient hotword detection through accelerometer," *In Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 301-315, 2015.

[27] Bardella, P., *et al.*, "Optimal sampling frequency in recording of resistance training exercises," *Sports Biomechanics*, 16(1), 102-114, 2017.

[28] Small, S., *et al.*, "Impact of reduced sampling rate on accelerometer-based physical activity monitoring and machine learning activity classification," *Journal for the Measurement of Physical Behaviour*, 4(4), 298-310, 2021.

[29] Anjum, A. and Muhammad U. I. "Activity recognition using smartphone sensors," *In 2013 IEEE 10th consumer communications and networking conference (CCNC)*, pp. 914-919, 2013.

[30] Duan, L., *et al.*, "Human lower limb motion capture and recognition based on smartphones," *Sensors*, 22(14), 5273, 2022.

[31] Elstub, L. J., *et al.*, "Effect of pressure insole sampling frequency on insole-measured peak force accuracy during running," *Journal of Biomechanics*, 145, 111387, 2022.

[32] Figueroa, R.L., *et al.*, "Predicting sample size required for classification performance," *BMC medical informatics and decision making*, 12, pp.1-10, 2012.

[33] Berisha, V., *et al.*, "Digital medicine and the curse of dimensionality," *NPJ digital medicine*, 4(1), 153, 2021

[34] Džaja, D., et al., "Accelerometer-based algorithm for the segmentation and classification of repetitive human movements during workouts," Automatika, 64(2), 211-224, 2023.

[35] Pernek, I., et al., "Exercise repetition detection for resistance training based on smartphones," *Personal and ubiquitous computing*, 17, 771-782, 2013.