Journal for the Measurement of Physical Behaviour, 2021, 4, 102-110
https://doi.org/10.1123/jmpb.2020-0016

Human Kinetics
ORIGINAL RESEARCH

Application of Convolutional Neural Network Algorithms for Advancing Sedentary and Activity Bout Classification

Supun Nakandala, Marta M. Jankowska, Fatima Tuz-Zahra, and John Bellettiere
University of California San Diego

Jordan A. Carlson
Children's Mercy Hospital

Andrea Z. LaCroix and Sheri J. Hartman
University of California San Diego

Dori E. Rosenberg
Kaiser Permanente Washington Health Research Institute

Jingjing Zou, Arun Kumar, and Loki Natarajan
University of California San Diego

Background: Machine learning has been used for classification of physical behavior bouts from hip-worn accelerometers; however, this research has been limited due to the challenges of directly observing and coding human behavior "in the wild." Deep learning algorithms, such as convolutional neural networks (CNNs), may offer better representation of data than other machine learning algorithms without the need for engineered features and may be better suited to dealing with free-living data. The purpose of this study was to develop a modeling pipeline for evaluation of a CNN model on a free-living data set and compare CNN inputs and results with the commonly used machine learning random forest and logistic regression algorithms. **Method**: Twenty-eight free-living women wore an ActiGraph GT3X+ accelerometer on their right hip for 7 days. A concurrently worn thigh-mounted activPAL device captured ground truth activity labels. The authors evaluated logistic regression, random forest, and CNN models for classifying sitting, standing, and stepping bouts. The authors also assessed the benefit of performing feature engineering for this task. **Results**: The CNN classifier performed best (average balanced accuracy for bout classification of sitting, standing, and stepping was 84%) compared with the other methods (56% for logistic regression and 76% for random forest), even without performing any feature engineering. **Conclusion**: Using the recent advancements in deep neural networks, the authors showed that a CNN model can outperform other methods even without feature engineering. This has important implications for both the model's ability to deal with the complexity of free-living data and its potential transferability to new populations.

Keywords: ActiGraph, activity classification, activPAL, feature engineering, free living

Numerous studies have shown that sedentary behavior can be deleterious to human health. This research has demonstrated that even for individuals with moderate levels of physical activity, the overall amount and the pattern of sitting has been connected to health outcomes such as cardiovascular disease, diabetes, and cancer mortality (Bellettiere et al., 2019; Chang et al., 2020; Knaeps et al., 2018). Accurately quantifying sedentary behavior is the foundation of studying its relationship with health. Previously, many studies assessed sedentary behavior with self-reported questionnaires (Patterson et al., 2018). Due to the ubiquity of sitting behaviors, self-reporting of sedentary time is subject to high recall bias, leading to unreliable or inaccurate results in younger (Atkin et al., 2012) and older adults (LaMonte et al., 2019).

There is increasing interest among researchers and health care providers in objective methods for measuring sedentary time and patterns; such measurements have been most commonly achieved using hip-worn accelerometers. In a review of 46 studies of sedentary behavior using objective measurement methods, 34 utilized a hip-worn accelerometer; 31 out of these 34 used an ActiGraph device (Powell, Herring, Dowd, Donnelly, & Carson, 2018). Objective measurement of an adult's sedentary time from hip-worn accelerometers is most often quantified using a cut-point-based threshold of <100 counts/min that is applied to the vertical axis (Migueles et al., 2017), even on triaxial accelerometers. While this approach has good accuracy for measuring the total amount of time spent sedentary, it misclassifies standing without ambulation and vehicle sitting, and is inaccurate for measuring sit-to-stand transitions and other sitting pattern metrics (Barreira, Zderic, Schuna, Hamilton, & Tudor-Locke, 2015; Carlson et al., 2019).

The desire for more accurate measurement of free-living behavior has led to alternate data processing techniques, such as machine learning (ML) algorithms. Numerous studies from the

Nakandala and Kumar are with the Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. Jankowska is with the Qualcomm Institute/Calit2, University of California San Diego, La Jolla, CA, USA. Tuz-Zahra, Bellettiere, LaCroix, Hartman, Zou, and Natarajan are with The Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA, USA. Carlson is with the Center for Children's Healthy Lifestyles and Nutrition, Children's Mercy Hospital, Kansas City, MO, USA, and the Department of Pediatrics, Children's Mercy Hospital and University of Missouri Kansas City, Kansas City, MO, USA. Rosenberg is with the Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA. Jankowska (majankowska@ucsd.edu) is corresponding author.

computer science domains have demonstrated the utility of ML methods for successful human activity recognition from sensor and accelerometer data (Ramasamy Ramamurthy & Roy, 2018). In a recent review focused specifically on ML models for predicting type, class, and intensity of physical/sedentary activity domains using data acquired from a single body-fixed accelerometer, 62 studies were identified as using a variety of ML models, including artificial neural networks (32), support vector machines (18), random forest (RF) (12), decision trees (11), and logistic regression (LR) (7) (Farrahi, Niemelä, Kangas, Korpelainen, & Jämsä, 2019). Farrahi et al. noted that most of the studies included in the review trained ML models on laboratory or prescribed activity data sets, leading to high accuracy levels. However, once algorithms are applied to free-living populations, accuracy rates fall below 80% accuracy thresholds (Farrahi et al., 2019). Possible reasons for this include skew of behaviors of interest in natural settings (e.g., very few transitions or stepping behavior compared with sedentary time) and data exclusion from laboratory-based data sets of messy or nonclassifiable behaviors, which are often abundantly present in free-living data (Bastian et al., 2015; Sasaki et al., 2016). It is becoming increasingly apparent that in order for ML algorithms to become more generalizable, they will need to be trained and calibrated on free-living populations that can provide continuous and multiday data in order to adequately account for a diversity of behaviors as well as to obtain enough data to potentially balance class types (Keadle, Lyden, Strath, Staudenmayer, & Freedson, 2019; Kerr et al., 2018).

Research on the application of deep learning methods to human activity recognition is showing promising results in computer science domains, which may translate into improving the flexibility and generalizability of activity classification (Nweke, Teh, Al-Garadi, & Alo, 2018; Wang, Chen, Hao, Peng, & Hu, 2019). Two main types of deep models have been applied in activity recognition: convolutional neural networks (CNNs) (Krizhevsky, Sutskever, & Hinton, 2012) and long short-term memory models (Guan & Plötz, 2017). The hallmark feature of deep neural network models is their ability to learn relevant features without relying on hand-engineered features (e.g., researcher processed features, such as mean vector magnitude), which can take considerable data processing and development time, potentially introduce bias, and make the generalizability of an algorithm to new populations challenging. The ability to learn relevant features is particularly useful for accelerometer data in free-living settings (relative to laboratory settings) due to the variability and complexity of behavior during free-living. CNNs have been shown to excel at adapting to new data sets, opening up possibilities for reducing the need for ML models to be trained for each new cohort or context (Rokni, Nourollahi, & Ghasemzadeh, 2018; Saeedi, Norgaard, & Gebremedhin, 2018). These aspects of CNNs make them a good ML candidate for identifying physical behaviors in free-living populations as well as offering flexibility in transferring developed CNNs from one population to another.

The objective of this study was to develop a modeling pipeline to evaluate a CNN model on a free-living data set of 28 individuals and compare CNN results with the commonly used ML RF and LR algorithms. We built off of previous work that developed an RF classifier for estimating sedentary time using several engineered features (Kerr et al., 2018), and in this study, we detailed considerations and steps for application of a CNN model to the same data set, comparing the use of engineered and raw features. Our goal was to differentiate sitting postures from upright postures (sitting, standing, and stepping) in raw triaxial accelerometer data obtained from hip-worn accelerometers during free-living for a 7-day period during waking hours. Ground truth activity labels were produced from a thigh-mounted activPAL device (PAL Technologies, Glasgow, Scotland, United Kingdom), which contains starting and ending events of standing, stepping, and sitting. ActivPAL has been shown to be a good measure of sitting time and of sit–stand transitions and has been used in previous studies for ground truth posture labeling (Barreira et al., 2015; Carlson et al., 2019; Kerr et al., 2018; Powell et al., 2018).

Methods

Data

Data for this study were collected from women (mean age = 62.7 years, *SD* = 7.3 years) enrolled in a cross-sectional study of sedentary behavior and breast cancer–related biomarkers among breast cancer survivors. Eligible participants were women diagnosed with Stages I–III breast cancer within the past 5 years who had completed active treatment (e.g., radiation, chemotherapy) and were fluent in English. Women were excluded if they had a primary or recurrent invasive cancer within the last 10 years (other than nonmelanomic skin cancer or carcinoma of the cervix in situ), were over 85 years of age, recently had bariatric surgery, were taking insulin or corticosteroid medications, or were diabetic (Hartman et al., 2018). All participants provided written informed consent, and ethical approval was obtained from the institutional review board of University of California, San Diego.

Data collection included two accelerometers (hip ActiGraph and thigh-worn activPAL). While all 30 participants had hip accelerometer data, two participants did not have thigh accelerometer data, and the final *n* for the study was 28. Participants wore an ActiGraph GT3X+ accelerometer device (ActiGraph LLC, Pensacola, FL) on a belt on the hip for 7 days during waking hours for an average of 854 min/day (*SD* of 46.7 min). Raw accelerometer data were collected in a time series format recorded at a 30-Hz frequency on three axes. Participants also wore the activPAL triaxial accelerometer (PAL Technologies Ltd., Glasgow, Scotland) on the anterior aspect of the thigh over the same 7-day period. Event files were output from activPAL software (version 7.2.32; PAL Technologies) as a time series with starting and ending times of sitting, standing, and stepping bouts. GT3X+ and activPAL data were time-stamp matched to create one output file at the resolution of 30 Hz for each user. The lower resolution activPAL data were repeated to match the higher resolution GT3X+ data, resulting in 9,239,038 s of concomitant activPAL and ActiGraph data across all participants and days. Periods of nonwear time greater than 60 s were identified using the Choi algorithm applied to the ActiGraph data (Choi, Ward, Schnelle, & Buchowski, 2012). Nonwear time was then removed from the combined (activPAL and ActiGraph) output file.

Exploratory Analysis—ActivPAL Data

Exploratory data analysis of the activPAL data was conducted prior to setting up ML models in order to assist with modeling decisions. Time distributions of the three activity types (sitting, standing, and stepping) were evaluated in aggregate (across participants and days) and distribution (Figure 1). Figure 1a shows a large skew between the three activity types. Sitting accounted for an average of 57% of total activity time, while stepping accounted for 13%. Box plots of the activity types (Figure 1b) showed large variations in the bout times of the activities. The median bout times were more than

2 min for sitting, 14 s for standing, and 7 s for stepping. Furthermore, 18% of stepping bouts and 16% of standing bouts were less than 3 s long. Figure 1 illustrates a cohort that did significantly more sitting than standing or stepping, often for longer periods that did not involve many transitions between sitting and standing.

Further data exploration was conducted by visualizing the raw accelerometer data. Figure 2 demonstrates eight random 5-s time windows of accelerometer instances for the three activPAL activity classes. Overall, variation in the accelerometer data was low for sitting and standing and high for stepping activity, as expected. However, outliers in these patterns were apparent, such as the fifth example of sitting, which has high variation and is unlikely to be a true sitting instance. Similarly, the second example of standing exhibits no variation and is unlikely to be a standing instance. These examples of discrepancies between the GT3X+ data and activPAL activity labels demonstrate the unique challenge for classifying activities in free-living compared with the laboratory data commonly used in the literature.

The exploratory data analysis provided information for decision points in the remaining analysis, which had to account for: (a) the highly skewed nature of the data toward sitting time, (b) high variation among bout times of different activities in which sitting occurred for longer periods than stepping or standing, (c) unreliability of very small bout times, and (d) the need for a filtering procedure to remove highly unlikely activPAL labels.

Prediction Time Window

The time window was used to extract windowed input from the time series data for feature engineering and to be fed into the ML algorithm. In the literature, different temporal window (or input context) sizes have been used, ranging from 1 to 60 s (Farrahi et al., 2019). In selecting a time window size, we considered activity bout times in the data set (which were small for stepping and standing bouts) and confidence in activPAL labels for very small bout times. More than 80% of activity bouts in the data set had bout times

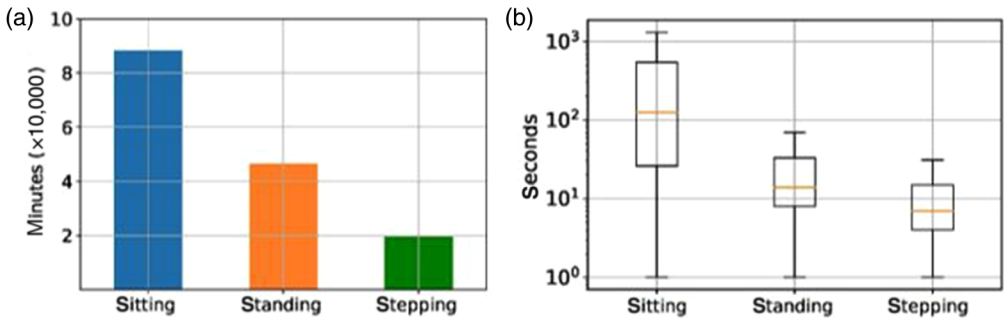

Figure 1 — (a) Mean time and (b) box plot distribution of bout time durations for the three activities (horizontal line in box is median level) aggregated across participants and days.

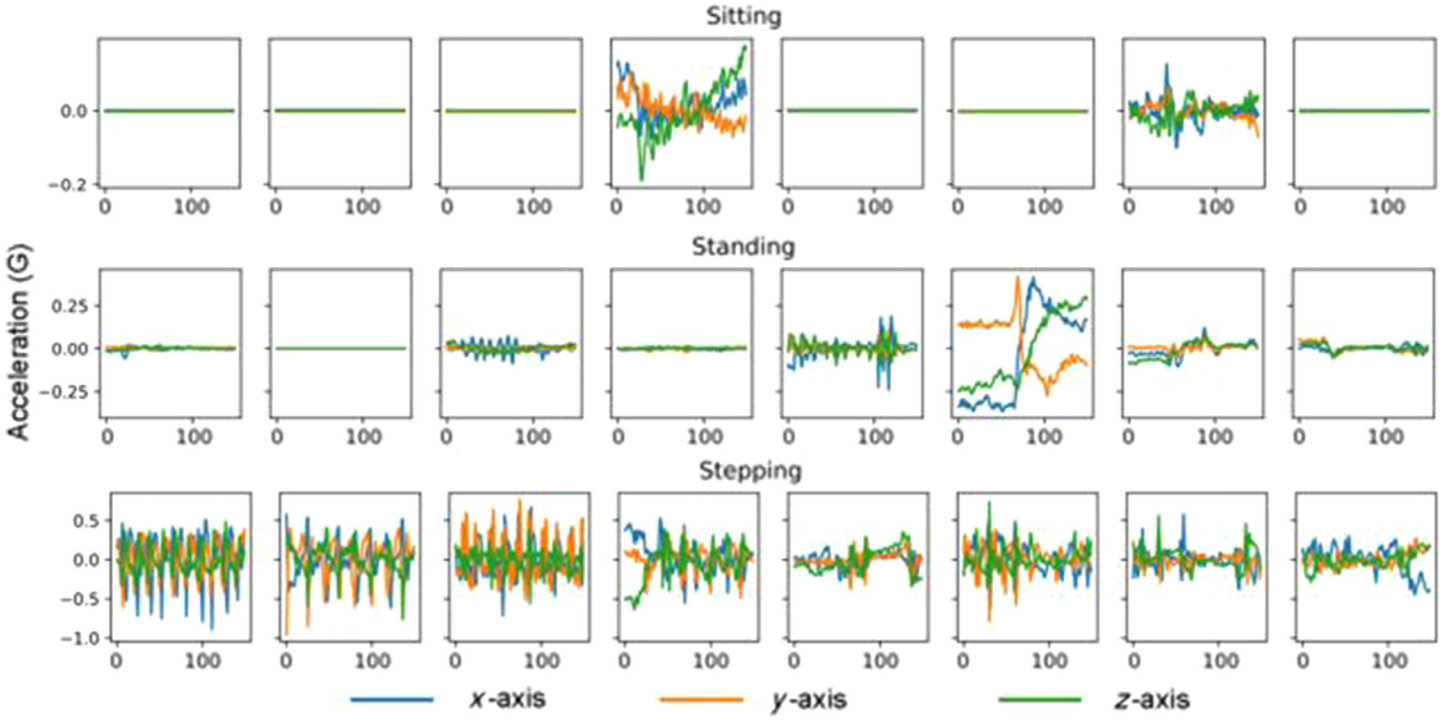

Figure 2 — Visualization of a random sample of GT3X+ 5-s windows in activPAL classified sitting, standing, and stepping bouts.

greater than 3 s; therefore, we selected a sliding window of 3 s ($3 \times 30 = 90$ data points) as our input context, which also served to reduce noise from bouts smaller than 3 s. The sliding window approach was applied to a continuous stream of input data, called segments, in order to extract the input contexts. Within these segments, the sliding window may map into regions in which the activPAL ground truth label is not the same; in other words, the window is entering into a transition with another label. Because the sliding window is small, we filtered out these border cases and considered only the time windows for which the ground truth label was consistent in the window. In total, 3.4% of the total data set was removed in filtering border cases.

Filtering

Filtering consisted of two steps: gravity removal and unlikely label removal. Participants were not instructed on which way (up or down) to wear the accelerometer, resulting in discrepancies in the orientation of the accelerometer between participants. Separating out the gravity component from the accelerometer signal helped to determine the orientation of the accelerometer device, which could be different for the same person for different moments and between different people. Gravity filtering was performed by applying a low-pass filter on the time series data as shown in the following algorithm, as notated using pseudo code. The algorithm took in an input accelerometer window $acc \in \mathbf{R}^{n*3}$, where n is the number of time steps in the window. For a 3-s input window, n is equal to 90 (30×3). Removing the gravity components transforms all axial acceleration components to the same scale and amplifies the local changes in the signal. Figure 3 shows examples of sitting, standing, and stepping activities before and after removing the gravity components from the accelerometer signal.

Algorithm: Remove Gravity Component

```
procedure RemoveGravity(acc)
    α = .9
    temp ←ZEROS(acc.shape)
```

(continued)

```
    temp[0, :] = (1 − α) * acc[0, :]
    for k ← 2 to acc.shape[0] do
        temp[k, :]←α × temp[k − 1, :] + (1 − α) * acc[k, :]
    gravity ←MEAN(temp, axis = 0)
    for k ← 1 to acc.shape[0] do
        acc[k, :]←acc[k, :] − gravity
    return acc, gravity
```

The second step in filtering removed unlikely activPAL labels for sitting, standing, and stepping, in line with the previously discussed visual inspection (e.g., Figure 2, Window 5 for sitting and Window 2 for standing). Wrong labels incorporate more noise into the ML models and hinder their learning and generalization capabilities. Therefore, we removed likely false labels using a simple heuristic. If a person is standing or stepping, the chances of the *SD* of their total acceleration $\left[v = \sqrt{\left(x^2 + y^2 + z^2\right)}\right]$ being close to zero is very low. Input time windows with labels corresponding to standing or stepping activities with an $SD \leq 10^{-4}$ were removed.

Feature Engineering

Increased discriminative power, reduction of noise, and removal of gravity can be achieved from feature engineering and extraction of triaxial accelerometer data. Removal of gravity helps to account for orientation mismatches between devices and acts as a data standardization step. In this study, we employed a feature engineering procedure for each 3-s window utilizing our group's previously developed and well-documented procedures, resulting in a total of 41 feature vectors (Ellis, Godbole, et al., 2014; Ellis, Kerr, et al., 2014; Kerr et al., 2016). Engineered features were divided into two main groups: time-domain features and frequency-domain features. Time-domain features included mean, *SD*, and percentiles, and frequency-domain features included entropy and power of certain frequencies, which were obtained by performing fast Fourier transform over the temporal window. All engineered features

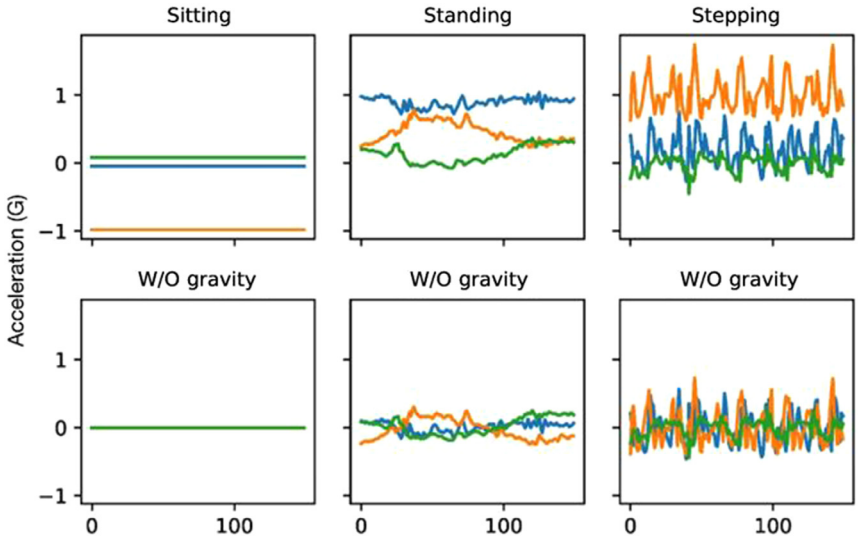

Figure 3 — Accelerometer data in 5-s windows for the postures sitting (left), standing (middle), and stepping (right) in original format (top row) and after removing the gravity component (bottom row).

are listed and described in Supplementary Table 1 (available online). Engineered features were standardized into a (−1, 1) scale.

ML Models

Three model sets were run (LR, RF, and CNN) for the purpose of classifying sitting, standing, and stepping activities. For all models, a train–validation–test split of 60–20–20 percentages was used. The split was based on individual users (16 users were selected as training data and six users each were selected as validation and test data) rather than sampling from the data pool of all users in order to allow for generalization to unseen users in the future. Models were trained to predict ground truth (activPAL) labels in independent 3-s intervals, thus ignoring the ordered time sequence and potential dependency across time intervals.

Hyperparameters are model-specific configurations that cannot be directly learned from the data, while parameters can be learned from the data. Hyperparameters are manually tuned until the model with the best prediction accuracy is found. The training data set is used to estimate the parameters of a model for prespecified hyperparameters, a validation set is used to find the best hyperparameters among the set of all hyperparameters used, and the test set is used to estimate the accuracy of the final model and hyperparameter selected. For tuning the hyperparameters in the models, the validation data set was used. Balanced accuracy rate (BAR) was chosen to be the performance metric for the study due to the significant activity class skew in the data set. BAR is the proportional average of accuracy in each category or class. Accuracy is the proportion of correctly classified instances out of the total. BAR is preferred over regular accuracy if the data are imbalanced. Results report BARs for training, validation, and test data sets for all the hyperparameters evaluated, and the test accuracy corresponding to the model that had the best validation accuracy was selected as the final accuracy metric for comparison. Classification reports were generated for the best model with precision, recall, and leave-one-out validation results. Precision is the proportion of correctly classified instances out of the predicted instances for each respective class. Recall is the proportion of correctly classified instances out of the actual or ground truth instances for each respective class. Leave-one-out cross-validation accuracy is estimated by training the model on all participants except one; the data of the participant who is left out are used as a test set. This process is repeated for all participants and the average accuracy of all the left-out participants is the leave-one-out cross-validation accuracy.

Our baseline model utilized an LR model, which was run using the scikit-learn (version 0.20.0) ML library in Python (version 2.7; Python Software Foundation, Fredericksburg, VA). Because the problem was a multiclass classification problem, we used multinomial logistic loss as the loss function. We used L2 regularization and tuned the model parameters using a validation data set. When feeding in raw data, triaxial accelerometer data in the 3-s window were flattened to produce a one-dimensional (1D) feature vector, which did not preserve the time series information. Flattening can be thought of as representing a matrix in row major order. The LR model was run on raw data (with and without removing the gravity component) and using engineered features.

The second model was an RF model, which was also run using scikit-learn. RF models are classifiers with high representational power compared with simpler linear models like LR. Thus, they have more learning and generalization capability. However, if not properly regularized, they tend to overfit for the training data due to their high representational power. The RF model used in this study had 100 trees and was regularized by setting the maximum depth of the trees, which we tuned by using a validation data set. The RF model was run on raw triaxial accelerometer data (with and without removing the gravity component) and using engineered features.

The third model was a CNN model. CNNs are a specialized form of neural network that excel at exploiting the spatial locality of information among features, such as relationships between neighboring pixels in images. Therefore, they have yielded near-human accuracy on benchmark image classification tasks, such as ImageNet (Krizhevsky et al., 2012). This same notion of locality of information can also be applied to the time domain, with temporal patterns resembling pixel variations. The dimensionality in this application is reduced from a two-dimensional (2D) spread of pixels to a 1D spread of time series values. Thus, such CNNs are also called 1D CNNs.

We trained a 1D CNN on raw accelerometer data as well as gravity component–removed accelerometer data. The CNN model had 6 layers, including convolution, pooling, dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), fully-connected, and softmax layers. The architecture of the CNN model is detailed in Supplementary Table 2 (available online). For dropout layers, a keep probability of .5 was used, and cross entropy loss was deployed as the loss function for training. To account for the significant class skew present in the data, we modified the cross entropy loss by multiplying it with values proportional to the class frequencies per the following equation, where $I_l(.)$ is the indicator function, and α_l is a value proportional to the class frequencies of label l.

$$\text{Loss} = \sum_{x \in X, y \in Y} \sum_{l \in \text{Labels}} \alpha_l \times l_l(y,\ l) \times \log[P(l == y|x)]$$

The CNN model was trained on the complete data set using back propagation and the Adam optimizer for 15 iterations (Kingma & Ba, 2015). The learning rate was tuned using a validation data set. Furthermore, we drew on recent work on wide and deep networks, which combined learning from deep neural networks and structured or engineered features, to generate a modified version of our CNN model called Wide-CNN (Cheng et al., 2016). Wide-CNN essentially augments the CNN model by concatenating the flattened Pooling5 features with the previously described hand-engineered features and adds two more fully connected layers. The architecture of the wide-CNN model is shown in Supplementary Figure 1 (available online). This model was trained like the original CNN model, and the learning rate was tuned using a validation data set. The final wide-CNN model and test dataset are available through the Github repository github.-com/ADALabUCSD/DeepPostures.

Experimental Setup

All experiments were run in a single node machine on CloudLab, a free and flexible cloud for research. The machine had two Intel Xeon Silver 4114 10-core CPUs at 2.20 GHz, 192 GB RAM (Intel, Santa Clara, CA), and one Nvidia P100 GPU (Nvidia, Santa Clara, CA). For training the LR and RF models, we used scikit-learn (version 0.20.0), and for training the CNN models, we used TensorFlow (version 1.9.0; Google, Mountain View, CA) with GPU support. Training of the RF model and feature engineering steps were parallelized to use all available cores in the machine. Runtimes for each experiment were as follows: LR model on raw features, 65 s; LR model on engineered features, 255 s; RF model on raw features, 292 s; RF model on engineered features, 202 s; CNN model on raw features, 1,035 s; CNN model on engineered features, 1,224 s; and feature engineering, 420 s.

Results

Ground Truth Activity Measures

On average, over the 7 days of wear time, participants engaged in 452.9 min of sitting per day, 231.7 min of standing, and 93.5 min of stepping. Participants wore devices for an average of 778.1 min per day.

Logistic Regression and RF Models

Results for the LR model are displayed in Table 1. Both the LR and RF models were evaluated with raw features, raw features after removing the gravity component, and engineered features. For tuning the L2 regularization factor, four different values were evaluated (0.1, 0.2, 1, and 10). The model with engineered features significantly outperformed the raw feature and gravity-removed feature models, with the highest BAR value at 0.76 for the L2 regularization value of 1. Results for the RF model are shown in Table 2. The tree depth was evaluated for four different values (20, 40, 60, and 80). For the RR model, BAR values across the three models were more similar than for the LR model, with all reaching values in the 0.7 range. The highest BAR (0.79) was achieved by the engineered features model with a tree depth of 20.

CNN Model

Results for the CNN and wide-CNN models are displayed in Table 3. The CNN model was evaluated with raw features and raw features with gravity removed, while the wide-CNN combed raw features with engineered features. Four learning rate values were evaluated for all models (0.01, 0.001, 0.0001, and 0.00001). All CNN models performed well, with most achieving BAR values higher than 0.8. The best performing model was the CNN with gravity removed features at a learning rate of 0.0001, with a BAR result of 0.84. Classification statistics were calculated for the CNN with gravity removed features, because it was the best performing model. Precision values for the three activities (sitting, standing, and stepping) were 0.93, 0.78, and 0.72, respectively, with an average precision value of 0.81. Recall values were 0.92, 0.74, and 0.86, respectively, with an average recall value of 0.84. F_1-Scores

Table 1 Logistic Regression Model Results

Raw features				Raw features without gravity				Engineered features			
L2 reg	Train acc.	Valid acc.	Test acc.	L2 reg	Train acc.	Valid acc.	Test acc.	L2 reg	Train acc.	Valid acc.	Test acc.
10	0.4651	0.4249	0.4689	10	0.4670	0.4251	0.4699	10	0.7481	0.7313	0.7543
1	0.4654	0.4243	0.4692	1	0.4669	0.4251	**0.4699**	1	0.7485	0.7318	**0.7550**
0.2	0.4657	0.4245	0.4692	0.2	0.4669	0.4251	0.4699	0.2	0.7485	0.7317	0.7549
0.1	0.4656	0.4245	**0.4699**	0.1	0.4669	0.4251	0.4699	0.1	0.7483	0.7317	0.7548

Note. All accuracy measures are BAR. BAR = balanced accuracy rates; acc = accuracy; reg = regularization. Bold values denote best test accuracy results across the four models.

Table 2 Random Forest Model Results

Raw features				Raw features without gravity				Engineered features			
Tree depth	Train acc.	Valid acc.	Test acc.	Tree depth	Train acc.	Valid acc.	Test acc.	Tree depth	Train acc.	Valid acc.	Test acc.
20	0.8035	0.7052	0.6941	20	0.8706	0.7442	0.7325	20	0.9220	0.7921	**0.7856**
40	0.9613	0.7462	0.7309	40	0.9991	0.7715	**0.7615**	40	0.9998	0.7895	0.7830
60	0.9959	0.7475	**0.7342**	60	0.9998	0.7713	0.7606	60	1.0000	0.7896	0.7822
80	0.9991	0.7472	0.7337	80	0.9999	0.7710	0.7607	80	1.0000	0.7895	0.7824

Note. All accuracy measures are BAR. acc = accuracy; BAR = balanced accuracy rates. Bold values denote best test accuracy results across the four models.

Table 3 CNN and Wide-CNN Model Results

CNN—Raw features				CNN—Raw features without gravity				Wide-CNN—Raw without gravity + engineered features			
Learn rate	Train acc.	Valid acc.	Test acc.	Learn rate	Train acc.	Valid acc.	Test acc.	Learn rate	Train acc.	Valid acc.	Test acc.
10^{-2}	0.7952	0.7390	0.7516	10^{-2}	0.7912	0.7376	0.7539	10^{-2}	0.7907	0.7814	0.7898
10^{-3}	0.8677	0.8411	**0.8336**	10^{-3}	0.8690	0.8425	0.8355	10^{-3}	0.8638	0.8245	0.8256
10^{-4}	0.8638	0.8401	0.8141	10^{-4}	0.8686	0.8495	**0.8406**	10^{-4}	0.8765	0.8235	**0.8350**
10^{-5}	0.8365	0.7963	0.8042	10^{-5}	0.8385	0.8262	0.8262	10^{-5}	0.8499	0.8215	0.8082

Note. All accuracy measures are BAR. BAR = balanced accuracy rates; CNN = convolutional neural network; acc = accuracy. Bold values denote best test accuracy results across the four models.

(harmonic mean of precision and recall) were 0.93, 0.76, and 0.78, respectively, with an average of 0.82. Finally, accuracy values were 0.92, 0.74, and 0.85, respectively, with an average value of 0.84. Confusion matrix results for prediction events (3-s windows) (Table 4) showed that sitting was misclassified as standing 15.7% of the time and as stepping at a rate of 3.1%. Standing was misclassified as sitting 6.7% of the time and as stepping at a rate of 31.0%. Stepping was misclassified as sitting 0.2% of the time and as standing at a rate of 4.6%. Leave-one-out cross-validation accuracy was performed on the CNN with gravity removed features model for each individual using a fixed learning rate of 0.0001. The maximum BAR was 0.93, the minimum was 0.67, and the average was 0.84.

Discussion

The results provide several insights into the task of identifying activities from accelerometer data using different types of ML models. The LR model on raw accelerometer features performed poorly, yielding a BAR of 0.47, which was marginally better than a trivial random baseline classifier accuracy of 0.33. Even removing the gravity component from the accelerometer data did not improve the prediction accuracy. However, when engineered features were introduced, the accuracy was boosted to 0.76, demonstrating the benefit of performing feature engineering. From these results, we can conclude that raw accelerometer features are not easily linearly separable, and the transformation performed by feature engineering makes the features more linearly separable. Results in Table 1 show that the accuracy of the LR model was not sensitive to the L2 regularization parameter. The training and test accuracies were also comparable, indicating that there was not much overfitting. It is also important to note that out of all the models, the LR is the most interpretable. Analysis of the corresponding coefficients of the learned model showed that the highest contributing feature for identifying sitting and stepping activities was the power of the 1 Hz frequency in the frequency domain. For standing, the highest contributing feature was the *SD* of the vector magnitude of the triaxial acceleration. For the RF models, the top three features were roll, pitch, and mean of vector magnitude of the triaxial acceleration.

Unlike the LR model, the RF model performed well even with the raw accelerometer features, yielding a balanced accuracy of 0.73. Removing the gravity component increased the accuracy by 3%, and using engineered features improved the accuracy up to 0.79. The relative better performance of the RF over the LR model can be attributed to the high representational power of the model. However, Table 2 shows how the RF model overfits to the training data, resulting in perfect classification for some tree depths. Regularization of the model by limiting the tree depth is important for improving generalization capability. However, the test accuracy dropped with regularization, suggesting that more labeled data would be needed to mitigate this overfitting.

Table 4 Confusion Matrix of Actual and Predicted Activities (per 3-s Window) Corresponding to the Best CNN Model Trained on Raw Features Without Gravity

	Predicted activity		
Actual activity	**Sitting**	**Standing**	**Stepping**
Sitting	3,39,223	26,617	1,881
Standing	24,722	1,25,625	19,109
Stepping	858	7,878	52,824

Note. CNN = convolutional neural network.

The CNN model outperformed the other two models even without using engineered features (raw accelerometer BAR = 0.83, raw accelerometer without gravity component BAR = 0.84). Augmenting the CNN model with engineered features using the wide-CNN architecture did not improve the accuracy. This could be because the CNN model already learned 19 relevant features automatically and including engineered features did not provide new information. Table 3 shows that the CNN models were sensitive to the learning rate parameter.

For training 15 iterations, a learning rate of 10^{-3} yielded the best accuracy for the CNN model with raw features and wide-CNN. A learning rate of 10^{-4} yielded the best accuracy for the CNN model with raw features without gravity. Learning rates that were too low or too high resulted in suboptimal final accuracies, showing the importance of hyperparameter tuning when training complex neural network ML models. Confusion matrix results showed that the CNN gravity removed model performed well for identifying sitting/stepping activities but moderately for identifying sitting activities. Precision, recall, and F1-score for sitting were high; however, standing and stepping activities had relatively low precision. Standing also had a low recall, while stepping had a better recall value. The model struggled most with identifying standing activities, which can be partially explained by Figure 2, in which standing raw accelerometer data at times looks like sitting and other times like stepping.

Because the sample size was small in this data set, it was computationally feasible to further evaluate the generalization capability of the CNN model by performing leave-one-out cross validation. The results for balanced test accuracies showed that most participants performed well, with an average BAR of 0.84. All participants performed over 0.75, except for one participant who had a BAR value of 0.67. However, note that the training accuracy was only at 0.86, unlike the RF. This suggests that amplifying the representation power of the CNN by making it deeper and larger could be beneficial, under the caveat that it may lead to more overfitting unless there is enough labeled data. We leave such exploration to future work.

Conclusions

This study evaluated the effectiveness of several ML methods, including CNNs, for the task of identifying activity classes from hip-worn accelerometer data in a free-living sample of 28 women. We employed the thigh-worn activPAL to specify the true labels, which has been shown to be similar to gold standard observations in previous studies (Steeves et al., 2015). However, hip or wrist accelerometry is still the most often utilized form of activity measurement in research studies. There is a need to improve not only accuracy/precision of activity classification from hip-worn accelerometer data but also the generalizability of generated models to other populations and contexts (Farrahi et al., 2019). Using the recent advancements in deep neural networks, we showed that a CNN model can outperform other methods, and, furthermore, it can do this without any feature engineering. The ability of these models to significantly reduce data processing time because of their ability to learn features from the data itself is a key advantage of CNNs over previously utilized machine learned models. Furthermore, these models have been shown to be highly generalizable to new populations (Rokni et al., 2018).

Ensuring accurate classification of free-living data with minimal feature engineering through researcher engagement can allow

for larger data sets to be analyzed, with better quantification of dose–response relationships between behaviors and health. An important next step in this research will be to independently validate the developed CNN model in a different population in order to test its generalizability. Another avenue of future research will be to apply combined CNN and long short-term memory models, which explicitly model the sequence information of the data. Previous research has shown that machine learning models have difficulty in identifying activity transitions, particularly in free-living data (Kerr et al., 2018). The application of a combined unstructured and structured ML model may be able to derive improvements for classification of activity transitions.

Limitations of the study include the small sample size. A larger sample will be especially important for assessing ML approaches in identifying transitions (such as sit to stand), because there tend to be much fewer occurrences of transitional behaviors in free-living populations compared with sitting, standing, and moving. Another limitation was a lack of sufficiently rich temporal features in the engineered data, which may contain useful information for predicting what behavior is most likely to be next within a sequence. In future studies, we will explore the utility of time as a feature in the models by combining CNN and long short-term memory models, which explicitly model longer-term temporal information in the data. A significant limitation of the current study was the exclusion of transitions. Algorithms for identifying behaviors in free-living populations must include identification of transitions from one behavior to another. Future development of the CNN model will focus on transition periods in order to allow for application in large free-living cohort studies.

Based on our findings in this free-living population, CNN models are a possible tool for dealing with the complexity of free-living data; however, future work focused on transitions is needed. Work in the computer science domain and even public health has relied to a large extent on laboratory or activity prescribed data sets. While these data offer clean examples of activities with messier transitions often removed, they may provide overly optimistic accuracy values for algorithms that then fall in accuracy statistics when applied to free-living data (Farrahi et al., 2019). This study provides compelling results for the ability of CNNs to adapt to free-living data.

Acknowledgments

This work was supported by grant number R01DK114945 from the National Institute of Diabetes and Digestive and Kidney Diseases. Research support for data collection was provided by pilot funding from the Department of Family Medicine and Public Health, UC San Diego. The work was also supported in part by a Hellman Fellowship, an NSF CAREER Award under award number 1942724, and a gift from VMware. The content is solely the responsibility of the authors and does not necessarily represent the views of any of these organizations. We thank the members of UC San Diego's Database Lab and Center for Networked Systems for their feedback on this work

References

Atkin, A.J., Gorely, T., Clemes, S.A., Yates, T., Edwardson, C., Brage, S., . . . Biddle, S.J.H. (2012). Methods of measurement in epidemiology: Sedentary behaviour. *International Journal of Epidemiology, 41*(5), 1460–1471. PubMed ID: 23045206 doi:10.1093/ije/dys118

Barreira, T.V., Zderic, T.W., Schuna, J.M., Hamilton, M.T., & Tudor-Locke, C. (2015). Free-living activity counts-derived breaks in sedentary time: Are they real transitions from sitting to standing? *Gait and Posture, 42*(1), 70–72. PubMed ID: 25953504 doi:10.1016/j.gaitpost.2015.04.008

Bastian, T., Maire, A., Dugas, J., Ataya, A., Villars, C., Gris, F., . . . Simon, C. (2015). Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: Laboratory-based calibrations are not enough. *Journal of Applied Physiology, 118*(6), 716–722. PubMed ID: 25593289 doi:10.1152/japplphysiol.01189.2013

Bellettiere, J., Healy, G.N., LaMonte, M.J., Kerr, J., Evenson, K.R., Rillamas-Sun, E., . . . LaCroix, A.Z. (2019). Sedentary behavior and prevalent diabetes in 6,166 older women: The objective physical activity and cardiovascular health study. *Journals of Gerontology—Series A Biological Sciences and Medical Sciences, 74*(3), 387–395. PubMed ID: 29726906 doi:10.1093/gerona/gly101

Carlson, J.A., Bellettiere, J., Kerr, J., Salmon, J., Timperio, A., Verswijveren, S.J.J.M., & Ridgers, N.D. (2019). Day-level sedentary pattern estimates derived from hip-worn accelerometer cut-points in 8–12-year-olds: Do they reflect postural transitions? *Journal of Sports Sciences, 37*(16), 1899–1909. PubMed ID: 31002287 doi:10.1080/02640414.2019.1605646

Chang, Y., Bellettiere, J., Godbole, S., Keshavarz, S., Maestas, J.P., Unkart, J.T., . . . Sears, D.D. (2020). Total sitting time and sitting pattern in postmenopausal women differ by Hispanic ethnicity and are associated with cardiometabolic risk biomarkers. *Journal of the American Heart Association, 9*(4), e013403. PubMed ID: 32063113 doi:10.1161/JAHA.119.013403

Cheng, H.T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., . . . Anil, R. (2016). Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems, ACM*. New York, NY :Association for Computing Machinery (pp. 7–10).

Choi, L., Ward, S.C., Schnelle, J.F., & Buchowski, M.S. (2012). Assessment of wear/nonwear time classification algorithms for triaxial accelerometer. *Medicine & Science in Sports & Exercise, 44*(10), 2009–2016. PubMed ID: 22525772 doi:10.1249/MSS.0b013e318258cb36

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J., & Kerr, J. (2014). Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. *Frontiers in Public Health, 2,* 36. PubMed ID: 24795875 doi:10.3389/fpubh.2014.00036

Ellis, K., Kerr, J., Godbole, S., Lanckriet, G., Wing, D., & Marshall, S. (2014). A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiological Measurement, 35*(11), 2191–2203. PubMed ID: 25340969 doi:10.1088/0967-3334/35/11/2191

Farrahi, V., Niemelä, M., Kangas, M., Korpelainen, R., & Jämsä, T. (2019). Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches. *Gait and Posture, 68,* 285–299. PubMed ID: 30579037 doi:10.1016/j.gaitpost.2018.12.003

Guan, Y., & Plötz, T. (2017). Ensembles of deep LSTM Learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(2), 1–28. doi:10.1145/3090076

Hartman, S.J., Marinac, C.R., Cadmus-Bertram, L., Kerr, J., Natarajan, L., Godbole, S., . . . Sears, D.D. (2018). Sedentary behaviors and biomarkers among breast cancer survivors. *Journal of Physical Activity and Health, 15*(1), 1–6. PubMed ID: 28682735 doi:10.1123/jpah.2017-0045

Keadle, S.K., Lyden, K.A., Strath, S.J., Staudenmayer, J.W., & Freedson, P.S. (2019). A framework to evaluate devices that assess physical

behavior. *Exercise and Sport Sciences Reviews, 47*(4), 206–214. PubMed ID: 31524786 doi:10.1249/JES.0000000000000206

Kerr, J., Carlson, J., Godbole, S., Cadmus-Bertram, L., Bellettiere, J., & Hartman, S. (2018). Improving hip-worn accelerometer estimates of sitting using machine learning methods. *Medicine & Science in Sports & Exercise, 50*(7), 1518–1524. PubMed ID: 29443824 doi:10.1249/MSS.0000000000001578

Kerr, J., Patterson, R.E., Ellis, K., Godbole, S., Johnson, E., Lanckriet, G., & Staudenmayer, J. (2016). Objective assessment of physical activity: Classifiers for public health. *Medicine & Science in Sports & Exercise, 48*(5), 951–957. PubMed ID: 27089222 doi:10.1249/MSS.0000000000000841

Kingma, D.P., & Ba, J.L. (2015, May 7–9). Adam: A method for stochastic gradient descent. ICLR: International Conference on Learning Representations, San Diego, CA.

Knaeps, S., Bourgois, J.G., Charlier, R., Mertens, E., Lefevre, J., & Wijndaele, K. (2018). Ten-year change in sedentary behaviour, moderate-to-vigorous physical activity, cardiorespiratory fitness and cardiometabolic risk: Independent associations and mediation analysis. *British Journal of Sports Medicine, 52*(16), 1063–1068. PubMed ID: 27491779 doi:10.1136/bjsports-2016-096083

Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*(2012), 1097–1105.

LaMonte, M.J., Lee, I.-M., Rillamas-Sun, E., Bellettiere, J., Evenson, K.R., Buchner, D.M., . . . LaCroix, A.Z. (2019). Comparison of questionnaire and device measures of physical activity and sedentary behavior in a multi-ethnic cohort of older women. *Journal for the Measurement of Physical Behaviour, 2*(2), 82–93. doi:10.1123/jmpb.2018-0057

Migueles, J.H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nyström, C., Mora-Gonzalez, J., Löf, M., . . . Ortega, F.B. (2017). Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Medicine, 47*(9), 1821–1845. PubMed ID: 28303543 doi:10.1007/s40279-017-0716-0

Nweke, H.F., Teh, Y.W., Al-Garadi, M.A., & Alo, U.R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications, 105,* 233–261. doi:10.1016/j.eswa.2018.03.056

Patterson, R., McNamara, E., Tainio, M., de Sá, T.H., Smith, A.D., Sharp, S.J., . . . Wijndaele, K. (2018). Sedentary behaviour and risk of all-cause, cardiovascular and cancer mortality, and incident type 2 diabetes: A systematic review and dose response meta-analysis. *European Journal of Epidemiology, 33*(9), 811–829. PubMed ID: 29589226 doi:10.1007/s10654-018-0380-1

Powell, C., Herring, M.P., Dowd, K.P., Donnelly, A.E., & Carson, B.P. (2018). The cross-sectional associations between objectively measured sedentary time and cardiometabolic health markers in adults—A systematic review with meta-analysis component. *Obesity Reviews, 19*(3), 381–395. PubMed ID: 29178252 doi:10.1111/obr.12642

Ramasamy Ramamurthy, S., & Roy, N. (2018). Recent trends in machine learning for human activity recognition—A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1254. doi:10.1002/widm.1254

Rokni, S.A., Nourollahi, M., & Ghasemzadeh, H. (2018, February 2–7). Personalized human activity recognition using convolutional neural networks. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, New Orleans, LA.

Saeedi, R., Norgaard, S., & Gebremedhin, A.H. (2018). A closed-loop deep learning architecture for robust activity recognition using wearable sensors. Proceedings—2017 IEEE International Conference on Big Data, Big Data 2017, Boston, MA, December 11–14, 2017. doi:10.1109/BigData.2017.8257960

Sasaki, J.E., Hickey, A.M., Staudenmayer, J.W., John, D., Kent, J.A., & Freedson, P.S. (2016). Performance of activity classification algorithms in free-living older adults. *Medicine & Science in Sports & Exercise, 48*(5), 941–950. PubMed ID: 26673129 doi:10.1249/MSS.0000000000000844

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929–1958.

Steeves, J.A., Bowles, H.R., McClain, J.J., Dodd, K.W., Brychta, R.J., Wang, J., & Chen, K.Y. (2015). Ability of thigh-worn ActiGraph and activPAL monitors to classify posture and motion. *Medicine & Science in Sports & Exercise, 47*(5), 952–959. PubMed ID: 25202847 doi:10.1249/MSS.0000000000000497

Wang, J., Chen, Y., Hao, S., Peng, X., & Hu, L. (2019). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters, 119,* 3–11. doi:10.1016/j.patrec.2018.02.010