HHS Public Access

Author manuscript

Stat Biosci. Author manuscript; available in PMC 2020 August 28.

Published in final edited form as:

Stat Biosci. 2019 ; 11(2): 334–354. doi:10.1007/s12561-019-09241-7.

Differentiating Between Walking and Stair Climbing Using Raw Accelerometry Data

William F. Fadel,
410 West 10th Street, Suite 3000, Indianapolis, IN 46202, Department of Biostatistics, School of Medicine & Richard M. Fairbanks School of Public Health, Indiana University

Jacek K. Urbanek,
2024 E. Monument Street, Suite 2-700, Baltimore, MD 21205, Division of Geriatric Medicine and Gerontology, Department of Medicine, School of Medicine, Johns Hopkins University

Steven R. Albertson,
723 W. Michigan St., SL280, Indianapolis, IN 46202, Department of Computer and Information Science, Indiana University-Purdue University Indianapolis

Xiaochun Li,
410 West 10th Street, Suite 3000, Indianapolis, IN 46202, Department of Biostatistics, School of Medicine & Richard M. Fairbanks School of Public Health, Indiana University

Andrea K. Chomistek,
1025 E. 7th Street, Bloomington, IN 47405, Department of Epidemiology and Biostatistics, Indiana University Bloomington

Jaroslaw Harezlak
1025 E. 7th Street, Suite C107, Bloomington, IN 47405, Department of Epidemiology and Biostatistics, Indiana University Bloomington

Abstract

Wearable accelerometers provide an objective measure of human physical activity. They record high frequency unlabeled three-dimensional time series data. We extract meaningful features from the raw accelerometry data and based on them develop and evaluate a classification method for the detection of walking and its sub-classes, i.e. level walking, descending stairs and ascending stairs. Our methodology is tested on a sample of 32 middle-aged subjects for whom we extracted features based on the Fourier and wavelet transforms. We build subject-specific and group-level classification models utilizing a tree-based methodology. We evaluate the effects of sensor location and tuning parameters on the classification accuracy of the tree models. In the group-level classification setting, we propose a robust feature inter-subject normalization and evaluate its performance compared to unnormalized data. The overall classification accuracy for the three activities at the subject-specific level was on average 87.6%, with the ankle-worn accelerometers showing the best performance with an average accuracy 90.5%. At the group-level, the average overall classification accuracy for the three activities using the normalized features was 80.2% compared to 72.3% for the unnormalized features. In summary, a framework is provided for better

Tel.: +1-317-274-2661, wffadel@iu.edu.

use and feature extraction from raw accelerometry data to differentiate among different walking modalities as well as considerations for study design.

Keywords

Classification Trees; Signal processing; Accelerometer; Physical activity; Walking

1 Introduction

The use of wearable accelerometers in public health research of physical activity (PA) has become increasingly popular. Unlike subjective methods, such as the widely used self-report questionnaires, wearable accelerometers offer a non-invasive objective measure of a person's PA. While subjective and objective methods may provide similar results with regard to qualitative findings for age and gender (e.g., males more active than females), the adherence to PA guidelines determined from accelerometers is substantially lower than from self-report [Troiano et al (2008)]. Furthermore, detailed quantification of PA attributable to specific activities is quite challenging and remains an elusive goal of PA monitoring research [Straczkiewicz et al (2016)]. Body acceleration is believed to be a valuable proxy for PA in the free-living environment. However, the usual method for describing accelerometer-measured PA is to use activity counts and a cut-point approach which classifies intensities of PA rather than the specific activity occurring [Veltink et al (1996);Esliger et al (2011);Zhang et al (2012);Straczkiewicz et al (2016)].

While use of accelerometers to assess PA may improve estimates for duration of time spent in activities of various intensities relative to self-report, the current methods may provide biased estimates of energy expenditures (EE). Activity counts are summarized over a given window, and then, they are compared to preset thresholds to determine whether a subject was engaged in sedentary, light, moderate, or vigorous PA. These methods are unable to differentiate between activities that produce similar total acceleration over time but that have differing EE [Pober et al (2006)]. For example, walking on a level surface and ascending stairs may produce similar levels of total acceleration, but the EE from ascending stairs is nearly double that of walking on a level surface [Campbell et al (2002)]. In fact, the relative metabolic rate of ascending stairs can be nearly five times that of walking on level ground depending on the speed of walking [Ohtaki et al (2005)]. Therefore, even short bouts of stair climbing can be an important distinction when considering an individual's overall EE throughout a given day. Although these cut-point methods are primarily used to summarize the raw accelerometry data, information about the structure of the data which may be pivotal to differentiating between activities is lost [Mannini et al (2013)]. Recent literature has attempted to address this problem using a signal processing approach. The Fast Fourier Transform (FFT) and discrete wavelet transform (DWT) have previously been used to develop more detailed feature sets for classification of different activity types [Zhang et al (2012)]. One disadvantage of the FFT is that information is lost from the time domain. The DWT addresses this problem by providing information in both the time and frequency domains, but due to the high dimensionality of the raw accelerometry data structure, implementing a windowing approach is still an attractive option. The short-time Fourier

transform (STFT) can then be implemented within a localized window recapturing the time information. However, this approach requires the choice of an appropriate window size be made Preece et al (2009);Urbanek et al (2018).

The windowing approach to data segmentation is common throughout the accelerometry literature. It has been demonstrated that smaller windows provide faster activity detection and computing time, but larger windows tend to perform better in the recognition of more complex activities [Banos et al (2014)]. There are no clear-cut rules when it comes to choosing window length, but it is important to consider the application prior to making a choice as some shorter activities could be obscured by noise in larger windows and longer activities may not be fully captured in shorter windows. Banos et al (2014) attempted to address this problem with an extensive study of the impact of window length on activity recognition. Although they conclude a window size of 1-2s provides the best trade-off between recognition speed and accuracy, their feature set consisted only of simple metrics such as mean, standard deviation, minimum, maximum, and mean crossing rate. When the interest lies in differentiation among similar activities such as walking and stair climbing, more detailed features must be implemented which require larger window sizes for higher resolution of spectral features.

In this paper, we describe the Indiana University Walking and Driving Study (IUWDS) that was designed to collect accelerometry data for walking, stair climbing, and driving in a simulated free-living environment. The study consisted of two separate trials, a walking trial and a driving trial. Figure 1 displays the raw accelerometry data from a single participant during the walking trial. Each subject was asked to complete five periods of walking on level ground and six periods, each, of ascending and descending stairs. All participants were instructed to perform each task at their usual pace to simulate data collected in a free-living environment. Using the complete data from both walking and driving trials, we were able to show that we can accurately differentiate between walking activities and driving with high accuracy [Straczkiewicz et al (2016)]. Therefore, the focus of this paper is on differentiating between the three walking modalities. Prior to any modelling, pre-processing steps were undertaken to extract meaningful information from the raw triaxial accelerometry data. Using a windowing approach, we extract features of the data from both the time and frequency domains. Most of the chosen features provide either a measure of the energy exerted from certain activities or measures of periodicity from the signal, and half of the features were derived from the FFT and DWT. Finally, extracted features are used to build classification trees at both the subject and population level. The classification tree was chosen because it has been shown to provide good classification of PA types [Bao and Intille (2004);Kwapisz et al (2011);Zhang et al (2012); Ellis et al (2016)]. Classification trees also provide an interpretable model that can be used to inform subsequent association studies as to which relevant features may be useful in modelling certain health related outcomes. The classification models were built under varying combinations of sensor location and window length. Model evaluation was performed to assess the impact of sensor location and window length on the classification accuracy for each of the three walking modalities.

The rest of this paper is organized as follows. In Section 2, we describe the data collection and labelling methods for the raw accelerometry data. In Section 3, we describe the signal

processing used to extract relevant features from the raw data. In Section 4, we describe the classification model and subsequent statistical models used to evaluate the properties of the classification models. In Section 6, we describe the results of classification and the impact of differing window sizes and sensor location on those results, and we also describe the features found to be most important for differentiation of the walking modalities. In Section 7, we provide a brief description of the study results and future research.

2 Data collection

Thirty-two adults (13 men, 19 women) participated in the IUWDS study. The data collected was used to identify patterns of walking, stair climbing, and driving from raw accelerometry data. The study was approved by the Institutional Review Board of Indiana University; all participants provided written informed consent. Participants wore four ActiGraph GT3X+ accelerometers: one on the left ankle, one on the right ankle, one on the left hip, and one on the left wrist. All four devices were synchronized to the same external clock providing parallel measurement at a sampling frequency of $100 Hz$ (i.e., 100 observations per second) for the four body locations. Each device was attached using velcro bands. The ankle accelerometers were worn on the outside of the ankles. The wrist accelerometer was worn similar to a regular watch on the top of the left wrist. The hip accelerometer was attached to the belt of the participant on the left hip, but when a belt was not available, the device was either attached to the corresponding belt loop or clipped to the waistband. Data were downloaded immediately following each participant's session. A human observer recorded the starting and stopping times for the walking study. All devices were initialized and data were downloaded using the manufacturer's software (ActiLife version 6.12.0) [http://actigraphcorp.com]. Table 1 contains demographic information for the study participants. Thirty-one of the participants were right handed while the remaining individual identified himself as ambidextrous. The study included a walking trial (approximately 0.66 miles) followed by a driving trial (approximately 12.8 miles). The walking trial included walking on level ground, ascending stairs, and descending stairs. Immediately after the walking period, participants were accompanied to their vehicle, and they then drove on a predefined route that included both highway and city driving. The walking trial lasted between 9.0 and 13.5 minutes while the driving trial lasted between 18 and 30 minutes, depending on traffic.

The data collection protocol requested participants to walk at their usual pace along a predefined course to simulate free-living activities. Our prior experience has demonstrated the inaccuracy of human observers labelling activities. In order to ensure accuracy of the starting and stopping times for different activities, participants were asked to clap three times at the beginning and end of each activity internally marking the raw accelerometry data for the wrist with three consecutive spikes in the signal. Using these internal markings within the data, we were able to accurately assign activity labels for each section of the protocol. Once the activity labels were assigned, the clapping signal ±0.5 second of data were deleted to mimic smooth transitions between activities. The walking trial consisted of five periods of walking on level ground, six periods of descending stairs, and six periods of ascending stairs. The data from one participant included an additional period of walking on level ground due to the participant briefly forgetting the instructions before turning around to

ascend the stairs. For the purposes of this paper, we focus strictly on data from the walking trial collected at the four sensor locations.

3 Signal processing

For each participant, we assume that we can identify periods of walking by utilizing the algorithm developed by Urbanek et al (2018) so we select only the walking trial data. Their method uses a frequency analysis approach to detect periodic activity within windowed portions of the raw triaxial accelerometry signal. They compute a ratio of the area of interest to the total area under the spectrum obtained by FFT that indicates periodic activity when this ratio exceeds a pre-specified threshold. This ratio (*ratio.VM*) is described in more detail below.

We consider the triaxial signal $\mathbf{x}(t) = \{x(t), y(t), z(t)\}$ where $x(t)$, $y(t)$, and $z(t)$ are the measurements along the three orthogonal axes of the device at time **t**. Participants walking while swinging their arm change the orientation of the wrist worn device with respect to earth's gravity which directly affects the measurement in each axis [Bai et al (2012); He et al (2014); Xiao et al (2016); Straczkiewicz et al (2016)]. In order to remove the effects of sensor orientation, we consider the vector magnitude, *VM*, where the vector magnitude at time t is defined as:

$$vm(t) = \sqrt{x(t)^2 + y(t)^2 + z(t)^2} \tag{1}$$

For feature extraction, we then use a sliding window approach to divide the signal into windows of 2.56, 5.12, and 10.24 seconds providing 256, 512, and 1024 samples per window (i.e. $2.56s \times 100Hz = 256$ samples), respectively. We use a set of windows of varying lengths in order to evaluate the impact of window size on feature importance and classification accuracy. Window sizes were chosen to ensure the number of samples in each window was a power of 2 to simplify computation of FFT and DWT and avoid the need for zero-padding. In addition, the smallest window of 2.56s ensures that a gait cycle is repeated at least twice. The number of windows analyzed varies by subject due to variability in the lengths of time to complete the walking trial. Similar to Zhang et al (2012), we extract features in both the frequency and time domains. The frequency domain features are derived from the FFT and the DWT of the *VM*. The thirteen features used are summarized in 2 and described in more detail in the following paragraphs. The sliding window FFT is referred to as the short-time Fourier Transform (STFT) [Sejdić et al (2009), Urbanek et al (2018), Straczkiewicz et al (2016)]. For a window of size τ, centered at time t, the STFT of the signal $vm(t)$ is defined as

$$VM(f,t) = \sum_{u=[t-\tau/2]}^{[t+\tau/2]} vm(u)h(u)e^{-i2\pi fu/\tau} \tag{2}$$

where f is the frequency index and the weights $h(u)$ assign more weight to observations close to t. We use the weights defined by the Hanning window, $h(u; \tau) = 0.5[1 - cos\{2\tau u/(\tau - 1)\}]$, as they have been shown to reduce aliasing, or a blurring of the spectrum [Harris

(1978); Urbanek et al (2018)]. The features extracted from the frequency spectrum of each window include: $f1$, $ratio.VM$, $p1$, and $p1.TP$.

Figure 2 provides a visual description of the features extracted from the FFT. While $ratio.VM$ and $p1.TP$ appear similar in concept, $p1.TP$ contrasts the power of each step versus the entire spectrum, while $ratio.VM$ contrasts multiple characteristics of walking versus the non-walking related portions of the spectrum. In essence, if we consider all relevant human movement to occur between 0.3-12.5 Hz, $p1.TP$ is measuring the energy associated with the step component of walking versus all other movements within a given window. In contrast, $ratio.VM$ is measuring the periodic content relative to the non-periodic content associated with the VM signal.

Additionally, we included two DWT features similar to Zhang et al (2012). The DWT of the signal $vm(t)$ is calculated from the wd() function in the R package **wavethresh**. The features extracted from the DWT of each window are given by the following equations:

$$DWT.VM2 = \sum_{j=\alpha}^{\beta} d_j^2 / VM^2 \tag{3}$$

$$DWT.TP = \sum_{j=\alpha}^{\beta} d_j^2 / \sum_{j=1}^{J} d_j^2 \tag{4}$$

where $d_j^2 = d_j^T d_j$ is the sum of squared DWT coefficient vector of VM at level j ($j = 1, \cdots, J$). In addition, VM^2 is the sum of the squared VM signal in each window. For our purposes, we selected α and β to cover the frequency range 0.78-6.25Hz, and J was selected to cover the frequency range 0-12.5Hz. When the noise in the signal is negligible, $DWT.VM2$ and $DWT.TP$ are nearly identical.

In addition to the FFT and DWT features, we included the vector magnitude count, VMC, which Urbanek et al (2018) and Straczkiewicz et al (2016) defined as

$$VMC(t) = 1/\tau \sum_{u=[t-\tau/2]}^{[t+\tau/2]} |vm(u) - 1/\tau \sum_{u=1}^{\tau} vm(u)| \tag{5}$$

where $VMC(t)$ is the VMC for the window of length τ centered at t and four features derived from the raw triaxial signal: activity intensity ($Act.Int = (s_x + s_y + s_z)/3$), $CORR.XY$, $CORR.XZ$, and $CORR.YZ$. We define $CORR.XY$, $CORR.XZ$, and $CORR.YZ$ as the Pearson correlation coefficient between the respective axes, and s_x, s_y, and s_z are the standard deviations of the x, y, and z axes of the accelerometry signal, respectively. The mean and standard deviation of the VM were included as the final two time domain features and defined as

$$Mean.VM(t) = \frac{1}{\tau} \sum_{u=[t-\tau/2]}^{[t+\tau/2]} vm(u) \tag{6}$$

and

$$SD.VM(t) = \sqrt{\frac{1}{\tau - 1} \sum_{u = [t - \tau/2]}^{[t + \tau/2]} [vm(u) - Mean.VM(t)]^2} \quad (7)$$

3.1 Feature normalization

Before fitting any population level classification model, it is important to normalize features at the subject level. As Xiao et al (2016) demonstrated, accelerometry data is not directly comparable across subjects. Figure 3 illustrates these subject to subject differences for the VM for a 10.24 second window of each walking activity for two subjects from our study. While the measured acceleration appear similar in nature, we can see that the magnitude of the signal for each activity is different across the two subjects. In addition, we also observe that the magnitude of the signal for descending stairs is the highest followed by level walking and then ascending stairs. Hence our motivation for normalization is to normalize all features to walking. The usual standardization simply centers data around the mean of the distribution and scales by the overall standard deviation. A reasonable assumption that we make is that level walking is the overwhelmingly dominant type of walking for the vast majority of human physical activity. Therefore, we employ a simple, yet novel normalization scheme of centering each feature around the median value and scaling by the median absolute deviation (MAD). For a feature w, calculate a pseudo z-score as

$$z^* = \frac{w - median(w)}{MAD(w)} \quad (8)$$

where $MAD(w) = 1.4826 * median|w_i - median(w)|$ and is calculated using a built in function in the R statistical software (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/mad.html). We make the assumption that level walking is the most common type of walking for everybody. In addition, we have observed that the magnitude of individuals' accelerometry signals is lowest for ascending stairs, followed by level walking, and then descending stairs. Combining these two assumptions, we can reasonably assume that the median for each of the features extracted (excluding the correlation features) would be representative of level walking. Standardizing the features in this way ensures that each z-score can be interpreted as a deviation from level walking.

4 Classification Model

All data management and modelling was performed using RStudio version 0.99.467 [RStudio Team (2015)]. Zhang et al (2012) showed that many machine learning algorithms provide satisfactory classification results, but the classification trees and support vector machine provide the best results. We chose classification trees for modeling the three types of walking activities due to their interpretability and ability to handle correlated predictors. We are interested in an interpretable model so that we can further understand what features are important for differentating between the three activities. This understanding of important features will help to inform subsequent statistical analyses of walking features with relation to health related outcomes. The classification tree methodology from the R package **rpart**

[Therneau et al (2015); Therneau and Atkinson (2015)] was used for the training and testing of our classification models. In both subject- and population-level classification, our data followed a similar structure where the response variable was *Activity* defined as a factor with three levels associated with walking on level ground, descending stairs, and ascending stairs. We use the thirteen features described in Section 3 as predictors in our models.

4.1 Subject-level classifier

We built a classification tree for each of our 32 subjects under the 12 combinations of window length (2.56, 5.12, and 10.24 seconds) and sensor location (left hip, left wrist, left ankle, and right ankle). In order to evaluate the performance of each classifier, cross-validation (CV) was implemented to investigate the classification accuracy of each model. To avoid over-training the classifier to identify a single activity (i.e., walking), we identified the activity with the fewest number of observations and chose 60% of that number, n_{min}, for the size of our training sets from each activity. All remaining observations were used for testing. This process was repeated 100 times for each subject under each scenario and the confusion matrix from the CV was used to evaluate the performance of each model. We fit a final tree for each participant using all data and assigning a uniform class prior to address the imbalance in the three activities.

4.2 Population-level classifier

The classification tree described in Section 4.1 was focussed on within-subject classification. Now, we will extend that methodology to the population level. For the population-level classifier, we considered the same 12 combinations of window length and sensor location as in Section 4.1, but we built a single model from all 32 subjects. Again, we used CV to evaluate the performance of the models, but in this case, we split our data into training sets consisting of all data from 20 randomly chosen subjects and tested on the remaining subjects. Each model was fit using uniform class priors to address the imbalance in the three activities as was done on the final subject-level models in the previous section. In addition, to avoid overtraining the classifier to the subjects used in the training sets, each model was pruned using the 1-SE rule [Therneau and Atkinson (2015)]. The model was then tested on the remaining subjects' data, and this process was repeated 100 times under each scenario. Final models were fit to all subjects' data using uniform class priors as before.

In addition to the models based on one sensor at a fixed window length, we evaluated the usefulness of combining information from the wrist and hip worn sensors to see if the combined information would improve classification accuracy. Due to concerns about compliance in larger studies, we only chose to combine wrist and hip worn sensors as these are most likely to encourage higher compliance over ankle worn sensors.

4.3 Model evaluation

The accuracy of each classification model was evaluated using the following metrics:

- $Sensitivity = Recall = True\ Positive\ Rate\ (TPR) = \frac{TP}{TP+FN}$
- $Specificity = True\ Negative\ Rate\ (TNR) = \frac{TN}{TN+FP}$

- $$Positive\ Predictive\ Value\ (PPV) = Precision = \frac{TP}{TP + FP}$$
- $$F1\ score = \frac{2 * PPV * Sensitivity}{PPV\ +\ Sensitivity}$$

True positives (*TP*) are defined as the number of windows in a given class that are correctly classified (e.g. classifying walking as walking). False positives (*FP*) are defined as the number of windows classified to given class, but they actually belong to a different class (e.g. classifying walking as descending stairs). True negatives (*TN*) are defined as the number of windows from a given class that are not classified as a different class (e.g. the number of windows for ascending and descending stairs that are not classified as walking). False negatives (*FN*) are defined as the number of windows in a given class that are classified to something else (e.g. the number of windows of walking that are classified to ascending or descending stairs). The above measures are defined for classification of one walking modality versus the other two.

In addition to classification accuracy, we evaluated the feature set to identify which predictors provided the best separation of the three walking activities. At each iteration, a ranking of variable importance was obtained and averaged across the 100 iterations per subject for the subject-level classifier. For the population-level classifier, each iteration represents an observation used for evaluation. The rankings range from 1 to 13, where 1 is the most important predictor and 13 is the least important.

For the subject-level classifiers, linear mixed models (LMM) were used to evaluate the effects of window size and sensor location on the classification accuracy for the three activities. Least-squares means were evaluated for multiple comparisons using a Tukey adjusted p-value.

5 Computational considerations

An important factor as to whether this method is scalable to larger studies is the time it takes to process the signal and train our model. We will consider average computing time for our study and scale these number up to the usual one to two week data collection. For the signal processing and feature extraction, the average computing time was around 35 seconds for the walking trial data. The average length of the walking trials were right around 11.5 minutes for males and females in the study. Therefore, we could reasonably expect the processing time for one week of data (10,080 minutes) collected at 100Hz to take (10,080/11.5) * 35 ≈ 30,678 seconds (or around 8.5 hours). Computing time was around 75 seconds to fit the population level model including training and testing with cross-validation. These models are computationally fast to fit, and we would not anticipate a drastic increase in the computational time with larger studies. Indeed, the added computation time would be at the signal processing level. All processing was performed in Windows 10 Enterprise on an Intel(R) Core(TM) i7-6700 CPU at 3.4GHz with 16GB of RAM on a 64-bit windows operating system.

6 Results

We applied the classification trees described in Section 4 to the walking trial data for the 32 participants in the IUWDS. As described in Section 2, data were collected from sensors placed at the left hip, left wrist, left ankle, and right ankle. Participants were instructed to walk at their usual pace along a predefined course that included walking on level ground, ascending stairs, and descending stairs. The clapping periods used to internally mark the beginning and end of each activity type were removed from the raw signal in order to mimic smooth transitions between activities. Prior to modelling, the raw data were preprocessed using the methods described in Section 3. Twelve classification trees were built for each participant and at the population level using the data collected from the 4 sensor locations and 3 window sizes (2.56s, 5.12s, and 10.24s). In addition, we built classification trees combining features extracted from left hip and left wrist for ech of the 3 window sizes. Training and testing data were constructed using the CV method described in Section 4. We evaluate each classifier in terms of sensitivity, specificity, PPV, and F1 score. Feature evaluation was performed to assess the average importance ranking of each feature included in the model.

6.1 Subject-level model evaluation

Figure 4 shows the results of the activity classification problem in terms of boxplots for the sensitivity, specificity, PPV and F1 score for all participants obtained from models built under each of the 12 window length and sensor location scenarios. We observed shorter window lengths and data collected at the wrist yield the lowest classification accuracy while larger windows and data collected at the ankles yield the highest classification accuracy. However, it appears from the top left panel of Figure 4 that there are differences in these trends for descending stairs. For descending stairs, the levels of sensitivity seem to be constant across window sizes for data collected from the left wrist and outperform the data collected from the hip. For the shorter window lengths (2.56 and 5.12 seconds), the sensitivity for the data collected from the wrist is higher than for the data collected at the hip.

We investigate the impact of sensor location and window length on the classification accuracy using LMMs. The main takeaways from the analyses showed that classification accuracy is highest when data are collected from the ankle worn sensors, but the hip and wrist worn sensors still provide useful information. Increasing the window size from 5.12 to 10.24 seconds provides only marginal improvements in the classification of level walking while the classification of stair climbing is best observed using the 5.12 second windows. More detailed results from these models can be found in Appendix A.

6.2 Population-level model evaluation

Figure 5 shows the sensitivity and specificity obtained through CV for the 12 classification models for both normalized and raw feature models. Because of the imbalance in the three activities, we observe very high sensitivity for walking, but we observe much lower sensitivity for ascending and descending stairs regardless of sensor location and window size. In addition, we see the normalized features outperform the raw features in nearly every

scenario. Figure 6 shows the PPV and F1 score for the 12 classification models for both the normalized and non-normalized feature models. Again, we notice the normalized features results are nearly always better than the results obtained using the raw features. Similar to what we observed at the subject-level, walking on level ground is the easiest activity to identify among the three types of walking, but it is also the most prevalent activity by a large margin. Instead, if we focus on the PPV in the top panel of Figure 6, we observe that the PPV is higher for left wrist versus left hip for descending stairs while the relationship is reversed (i.e., left hip higher than left wrist) for ascending stairs. The left and right ankles yield nearly identical results in terms of model performance.

When combining information from the hip- and wrist-worn sensors, the most notable benefits are in the accuracy of classifying ascending and descending stairs. When using only the hip- or wrist-worn sensors for classification, descending stairs was poorly identified for the hip-worn data (ranging from 56% to 66% accuracy for the 3 window sizes) and ascending stairs was poorly identified for the wrist-worn data (ranging from 56% to 63% accuracy for the 3 window sizes). When combining information from these two locations, the accuracy for descending stairs is between 70% and 74% and the accuracy for ascending stairs is between 67% and 71%. The most balanced classification using the combined information is for the 5.12 second window length. This is understandable since climbing stairs tends to be a shorter activity and larger window sizes may introduce noise that makes it difficult to differentiate between level walking and stair climbing.

6.3 Feature evaluation

Figure 7 shows the distributions of the feature importance rankings for the subject-level classifiers with differing window lengths and sensor locations. For features extracted from the wrist data, we see consistently across all window sizes, the top five most important features are *SD.VM, VMC, DWT.VM2, Act.Int*, and $p1$ implying features that measure changes in the intensity of the acceleration are best at differentiating between types of walking from wrist worn devices. The same five features are also ranked most important for the hip data with 2.56s windows. The hip data with 5.12s windows includes those same features in the top six important features but also include *Corr.XY* with a large amount of variability in importance between subjects. When data is collected at the hip and 10.24s windows are used, the most important feature becomes *ratio.VM* implying improved resolution of the FFT spectrum improves classification. The top two features most important for both the left and right ankle data with 2.56s windows are *Mean.VM* and $p1$. Consistently, $p1$ and $p1.TP$ appear in the top three most important features for the ankle data with 5.12s and 10.24s windows which implies the amplitude of $f1$ plays a significant role in differentiating between types of walking when data are collected from the ankle.

Figure 8 shows the feature importance for the 12 scenarios of window length and sensor location for the population-level classifiers. Similar to what we observed at the subject level, we see that for data collected from the left wrist, the most important features are those features which measure the variation in measured acceleration (i.e. *SD.VM, VMC, DWT.VM2, Act.Int*, and $p1$). For the data collected at the hip, the same five features are ranked the highest with exception that *ratio.VM* becomes the most important variable for

window lengths of 10.24 seconds. Again, this is most likely attributable to the need for higher resolution of the walking spectra before $ratio.VM$ can be accurately measured. Consistently, $p1$ and $p1.TP$ are ranked highly for the models built from ankle data. This is consistent with our previous findings at the subject level, and indicate the magnitude of the walking spectra at the dominant frequency is quite useful for differentiating between types of walking when data is collected at the ankle.

7 Discussion

We have proposed a classification tree-based method for differentiating between walking on level ground, ascending stairs, and descending stairs using accelerometry data. Relevant features were extracted from the raw data using a combination of frequency analysis features and time domain features and a range of window sizes (e.g., 2.56, 5.12, and 10.24 seconds). In Section 6.1, we showed that we can achieve very good classification results using the proposed methods for classification within subjects. In Section 6.2, we took a step forward in trying to build a population level classification model under a number of window size and sensor location combinations and proposed a novel normalization of features to standardize all activities to walking.

The within-subject methods described in Section 4.1 are more accurate, but in larger scale studies, it may not be feasible to obtain training data for every subject. The population-level models detailed in Section 4.2 serve as an important step towards our ultimate goal of building a reliable classification model. We showed that a novel, yet simple, normalization of the features can improve between subject classification results in nearly all scenarios and activities.

The data from the IUWDS was collected in a simulated free-living environment from relatively healthy adults ranging in age from 23 to 54 years. The large heterogeneity in the study population, with respect to age, BMI, and gender, enhances the generalizability of our results. A next step for this research will certainly include conducting similar analyses in smaller groups of more homogeneous individuals to assess the accuracy of additional population specific models. In addition to creating classifiers for more homogeneous groups, more sophisticated normalization techniques, or combinations of techniques, may improve the accuracy of the proposed models.

Acknowledgements

This paper was made possible, in part, with support from the Indiana Clinical and Translational Sciences Institute Design and Biostatistics Pilot Grant funded, in part by Grant UL1TR001108, from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award. Jaroslaw Harezlak has received funding from the National Institute of Mental Health research grant R01MH108467.

A: Supplemental results tables

Table 3:

LS means of sensitivity for sensor location by activity for the subject-level classifiers.

Sensor Location	Activity	Mean	Lower CL	Upper CL
Left Wrist	Ascending	0.844	0.829	0.858
Left Wrist	Walking	0.852	0.837	0.866
Left Hip	Descending	0.863	0.849	0.877
Left Hip	Ascending	0.869	0.855	0.883
Left Wrist	Descending	0.874	0.859	0.888
Right Ankle	Ascending	0.885	0.870	0.899
Left Ankle	Descending	0.888	0.874	0.903
Left Ankle	Ascending	0.889	0.875	0.904
Right Ankle	Descending	0.889	0.875	0.904
Left Hip	Walking	0.900	0.885	0.914
Left Ankle	Walking	0.938	0.924	0.953
Right Ankle	Walking	0.939	0.925	0.954

[1]Groups with similar numbers are not significantly different from each other.

Table 4:

LS means of sensitivity for window length by activity for the subject-level classifiers.

Window Length	Activity	Mean	Lower CL	Upper CL
2.56s	Ascending	0.855	0.841	0.868
2.56s	Descending	0.855	0.842	0.869
10.24s	Ascending	0.876	0.862	0.889
2.56s	Walking	0.879	0.865	0.892
5.12s	Ascending	0.885	0.871	0.898
5.12s	Descending	0.889	0.876	0.903
10.24s	Descending	0.891	0.878	0.905
5.12s	Walking	0.914	0.901	0.927
10.24s	Walking	0.929	0.916	0.942

Table 5:

LS means of overall classification accuracy for sensor location for the subject-level classifiers.

Sensor Location	Mean	Lower CL	Upper CL
Left Wrist	0.856	0.845	0.868
Left Hip	0.877	0.865	0.889
Right Ankle	0.904	0.893	0.916
Left Ankle	0.905	0.894	0.917

Table 6:

LS means of overall classification accuracy for window length for the subject-level classifiers.

Window Length	Mean	Lower CL	Upper CL
2.56s	0.863	0.852	0.874
5.12s	0.896	0.885	0.907
10.24s	0.899	0.887	0.910

References

Bai J, Goldsmith J, Caffo B, Glass TA, Crainiceanu CM (2012) Movelets: A dictionary of movement. Electronic journal of statistics 6:559 [PubMed: 23293708]

Banos O, Galvez JM, Damas M, Pomares H, Rojas I (2014) Window size impact in human activity recognition. Sensors 14(4):6474–6499 [PubMed: 24721766]

Bao L, Intille SS (2004) Activity recognition from user-annotated acceleration data In: International Conference on Pervasive Computing, Springer, pp 1–17

Campbell KL, Crocker P, McKenzie DC (2002) Field evaluation of energy expenditure in women using tritrac accelerometers. Medicine and science in sports and exercise 34(10):1667–1674 [PubMed: 12370570]

Ellis K, Kerr J, Godbole S, Staudenmayer J, Lanckriet G (2016) Hip and wrist accelerometer algorithms for free-living behavior classification. Medicine and science in sports and exercise 48(5):933–940 [PubMed: 26673126]

Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG (2011) Validation of the genea accelerometer. Med Sci Sports Exerc 43(6):1085–1093 [PubMed: 21088628]

Harris FJ (1978) On the use of windows for harmonic analysis with the discrete fourier transform. Proceedings of the IEEE 66(1):51–83

He B, Bai J, Zipunnikov VV, Koster A, Caserotti P, Lange-Maia B, Glynn NW, Harris TB, Crainiceanu CM (2014) Predicting human movement with multiple accelerometers using movelets. Medicine and science in sports and exercise 46(9):1859–1866 [PubMed: 25134005]

Kwapisz JR, Weiss GM, Moore SA (2011) Activity recognition using cell phone accelerometers. ACM SigKDD Explorations Newsletter 12(2):74–82

Mannini A, Intille SS, Rosenberger M, Sabatini AM, HaskellW(2013) Activity recognition using a single accelerometer placed at the wrist or ankle. Medicine and science in sports and exercise 45(11):2193 [PubMed: 23604069]

Ohtaki Y, Susumago M, Suzuki A, Sagawa K, Nagatomi R, Inooka H (2005) Automatic classification of ambulatory movements and evaluation of energy consumptions utilizing accelerometers and a barometer. Microsystem technologies 11(8-10):1034–1040

Pober DM, Staudenmayer J, Raphael C, Freedson PS, et al. (2006) Development of novel techniques to classify physical activity mode using accelerometers. Medicine and science in sports and exercise 38(9):1626 [PubMed: 16960524]

Preece SJ, Goulermas JY, Kenney LP, Howard D (2009) A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. IEEE Transactions on Biomedical Engineering 56(3):871–879 [PubMed: 19272902]

RStudio Team (2015) RStudio: Integrated Development Environment for R. RStudio, Inc, Boston, MA, URL http://www.rstudio.com/

Sejdić E, Djurović I, Jiang J (2009) Time–frequency feature representation using energy concentration: An overview of recent advances. Digital Signal Processing 19(1):153–183

Straczkiewicz M, Urbanek J, Fadel W, Crainiceanu C, Harezlak J (2016) Automatic car driving detection using raw accelerometry data. Physiological Measurement 37(10):1757 [PubMed: 27653528]

Therneau T, Atkinson B (2015) An introduction to recursive partitioning using the rpart routines. Mayo Foundation: Rochester, MN

Therneau T, Atkinson B, Ripley B (2015) rpart: Recursive Partitioning and Regression Trees. URL http://CRAN.R-project.org/package=rpart, r package version 4.1-10

Troiano RP, Berrigan D, Dodd KW, Masse LC, Tilert T, McDowell M, et al. (2008) Physical activity in the united states measured by accelerometer. Medicine and science in sports and exercise 40(1):181 [PubMed: 18091006]

Urbanek JK, Zipunnikov V, Harris T, Fadel W, Glynn N, Koster A, Caserotti P, Crainiceanu C, Harezlak J (2018) Prediction of sustained harmonic walking in the free-living environment using raw accelerometry data. Physiological measurement 39(2):02NT02

Veltink PH, Bussmann HJ, De Vries W, Martens WJ, Van Lummel RC (1996) Detection of static and dynamic activities using uniaxial accelerometers. IEEE Transactions on Rehabilitation Engineering 4(4):375–385 [PubMed: 8973963]

Xiao L, He B, Koster A, Caserotti P, Lange-Maia B, Glynn NW, Harris TB, Crainiceanu CM (2016) Movement prediction using accelerometers in a human population. Biometrics 72(2):513–524 [PubMed: 26288278]

Zhang S, Rowlands AV, Murray P, Hurst TL, et al. (2012) Physical activity classification using the genea wrist-worn accelerometer. PhD thesis, Lippincott Williams and Wilkins

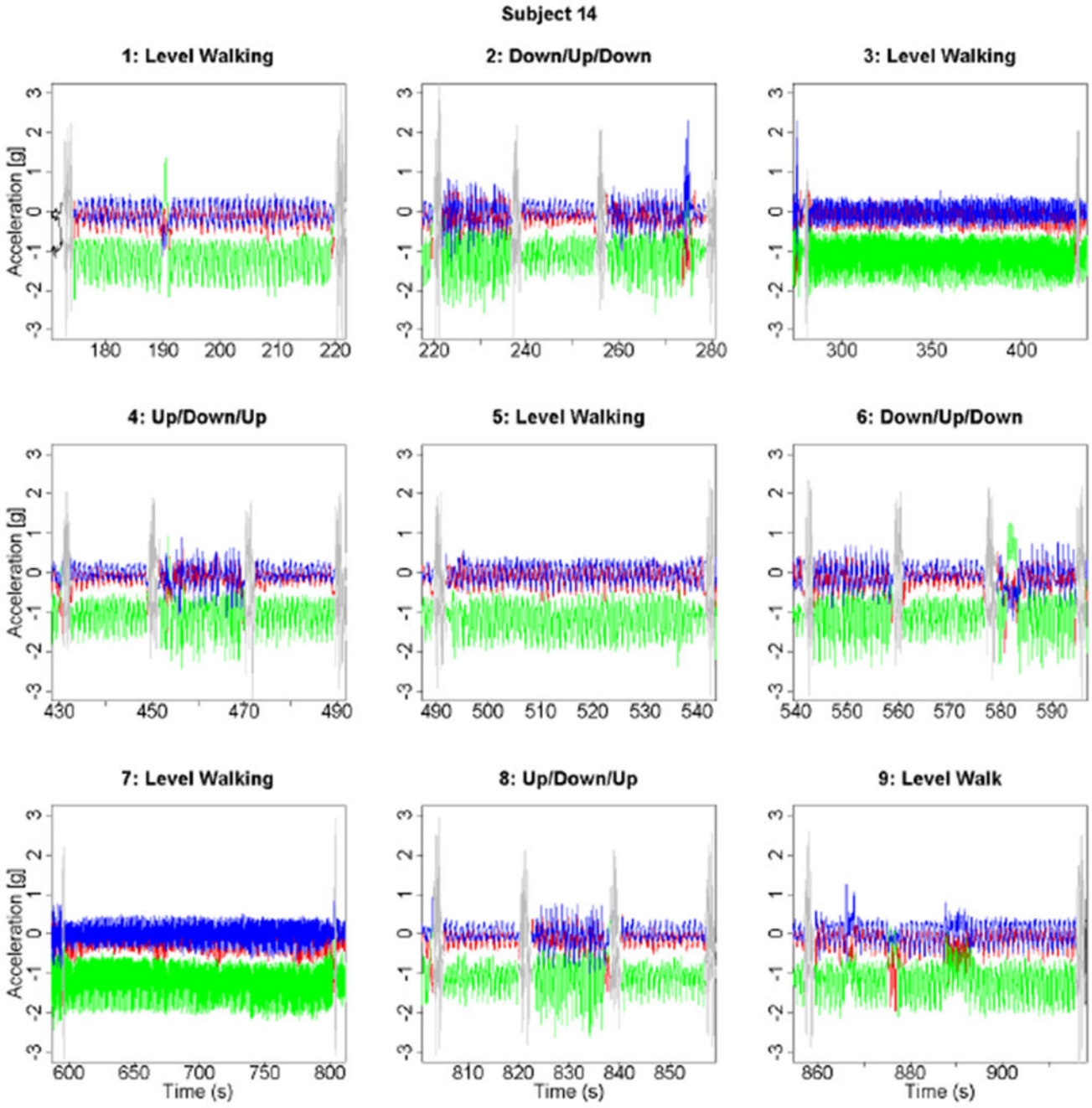

Fig. 1:

Triaxial raw accelerometry data for Subject 14 during the walking trial. Each panel represents different sections of the walking trial, and the red, blue, and green lines represent the acceleration measured from the three axes. The top left panel contains data from the first segment of walking on level ground from the start of the trial to the first set of stairs. The top middle panel represents the first set of stairs where the participant descended the stairs, ascended the stairs, and descended the stairs again prior to proceeding into walking on the second walking section (top right panel).

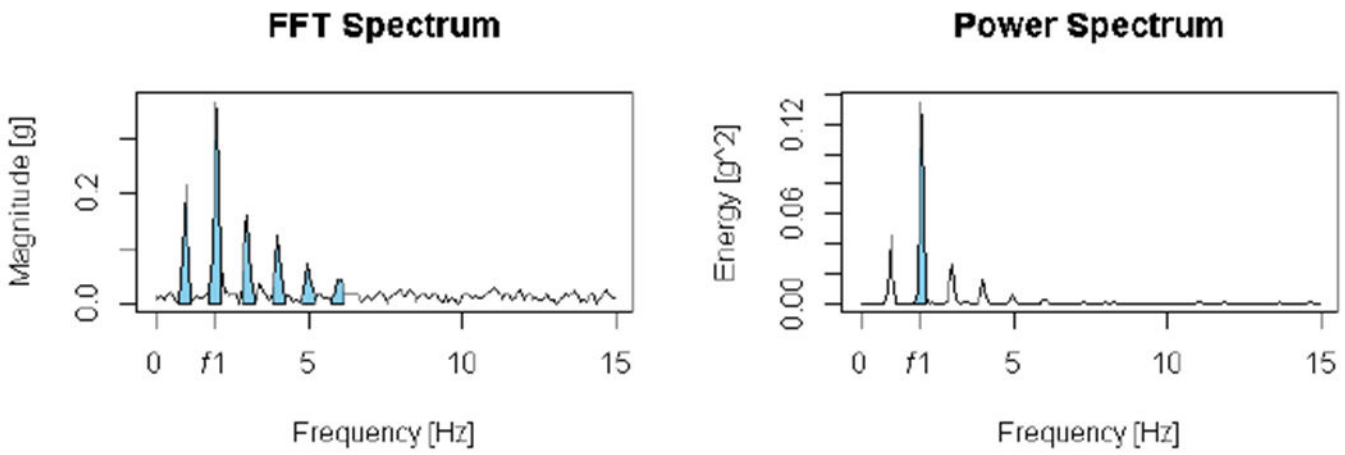

Fig. 2:

Fourier spectrum (left) and power spectrum (right) with shaded regions describing the features derived from the FFT. In the figure on the left, the shaded region represents the numerator of **ratio.VM**, and the dominant frequency is labeled as **f1**. In the figure on the right, the shaded region represents **p1**.

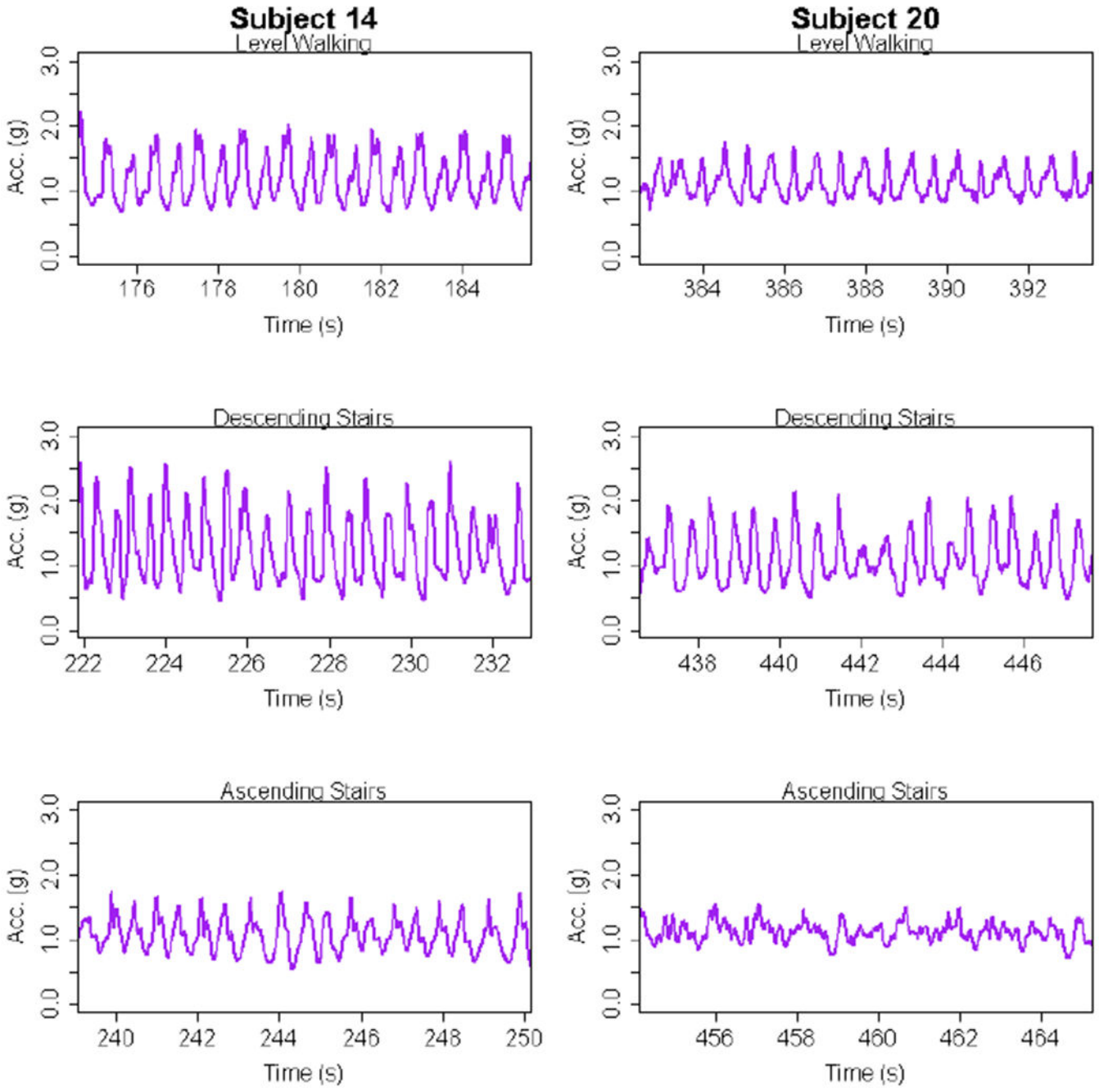

Fig. 3:

Vector magnitude for 10.24s windows of level walking (top row), descending stairs (middle row), and ascending stairs (bottom row) for Subject 14 (left column) and Subject 20 (right column).

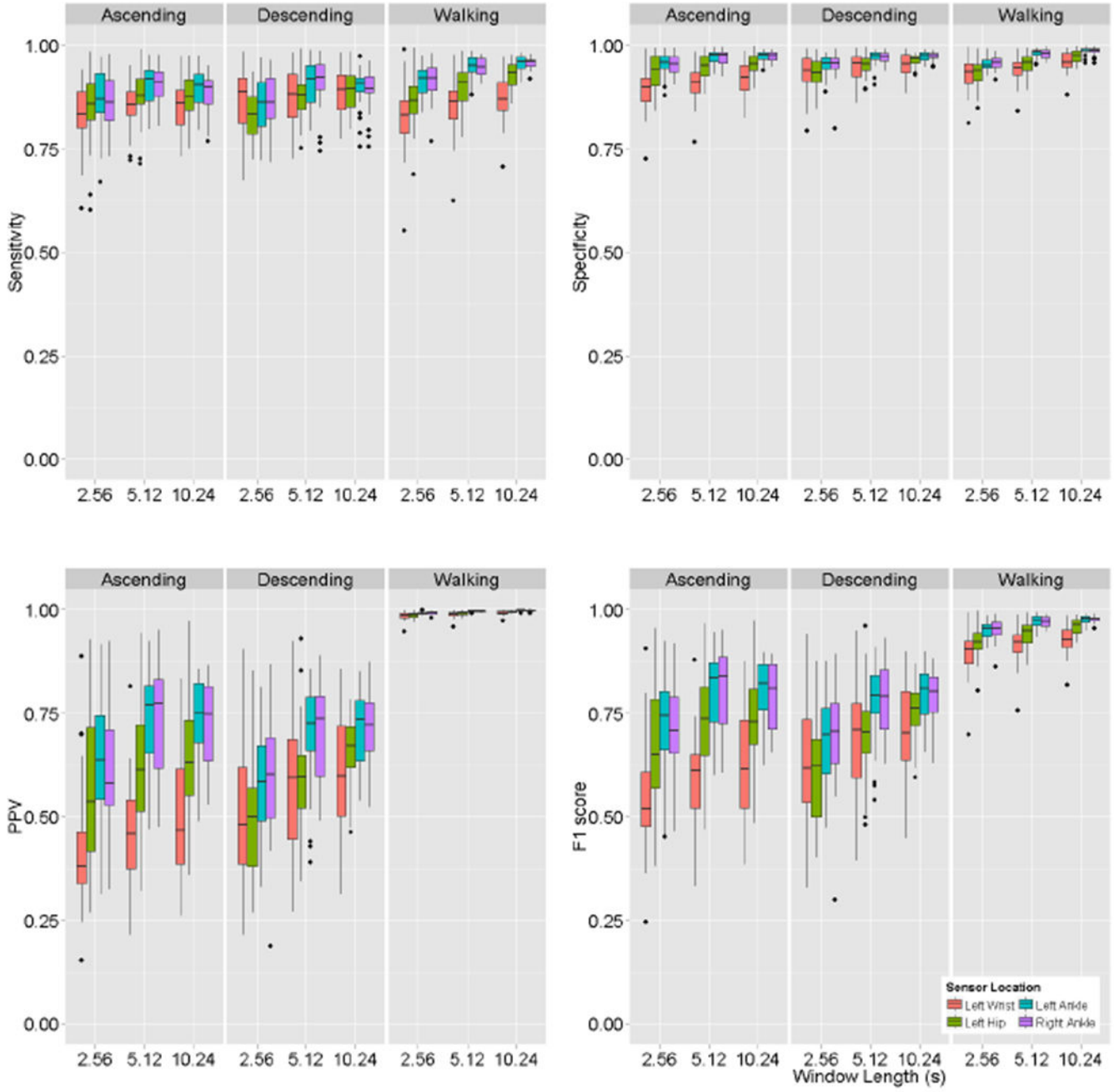

Fig. 4:

Boxplots for sensitivity (top left), specificity (top right), PPV (bottom left), and F1 score (bottom right) across participants by activity, sensor location, and window length for the subject-level classifiers.

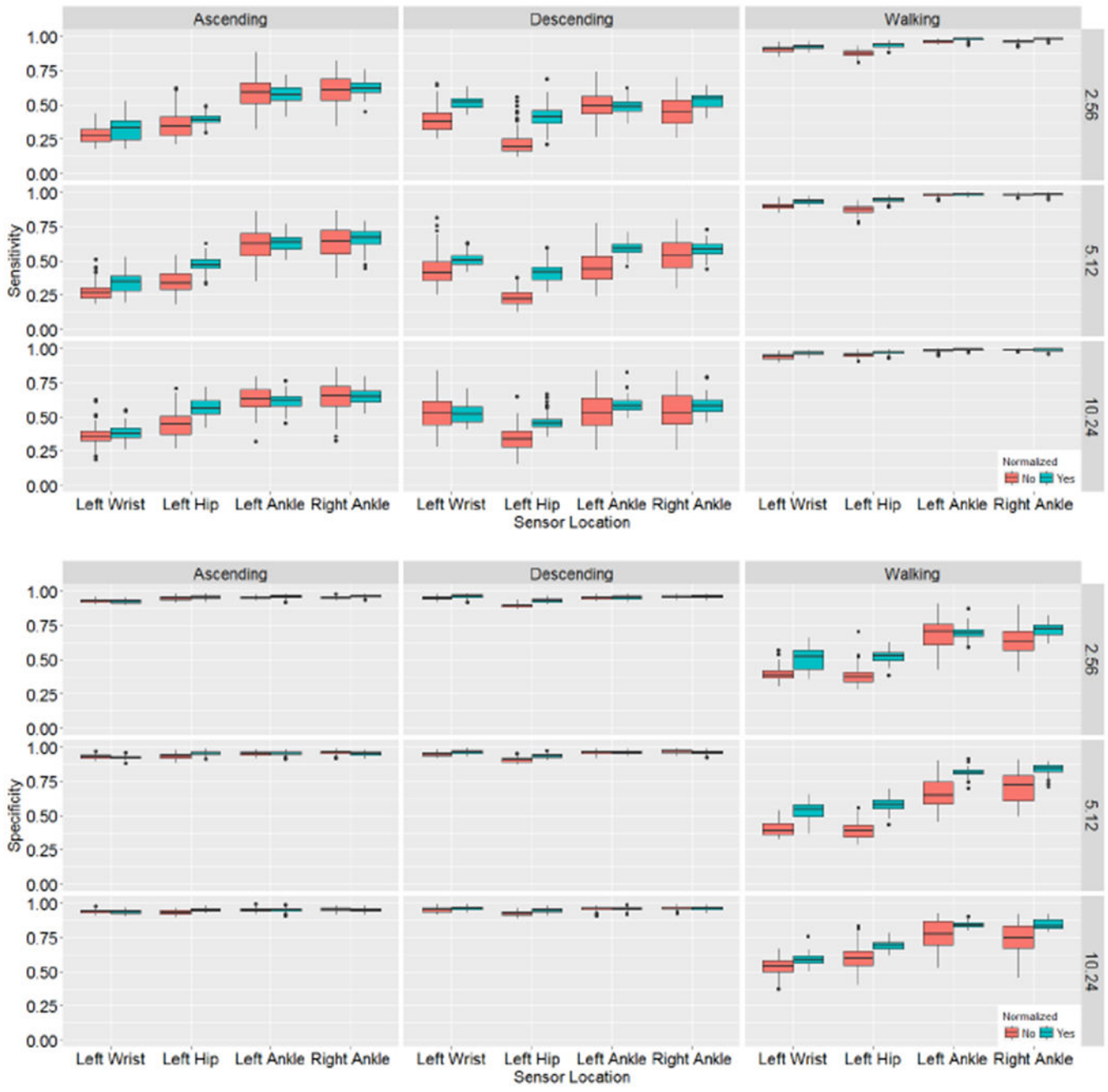

Fig. 5:

Sensitivity and specificity by activity, sensor location, and window length for the population-level classifiers.

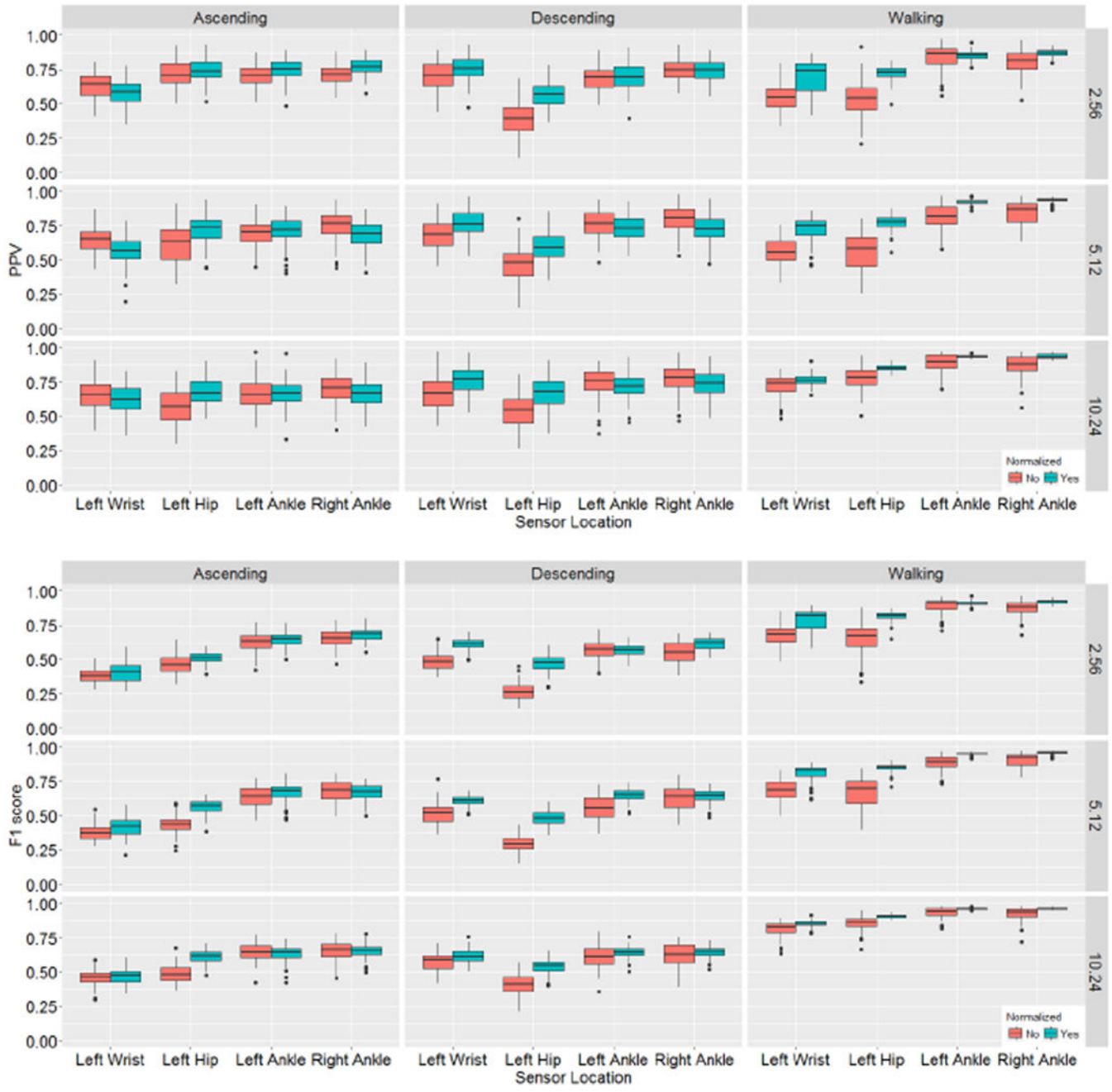

Fig. 6:

Positive predictive value and F1 score by activity, sensor location, and window length for the population-level classifiers.

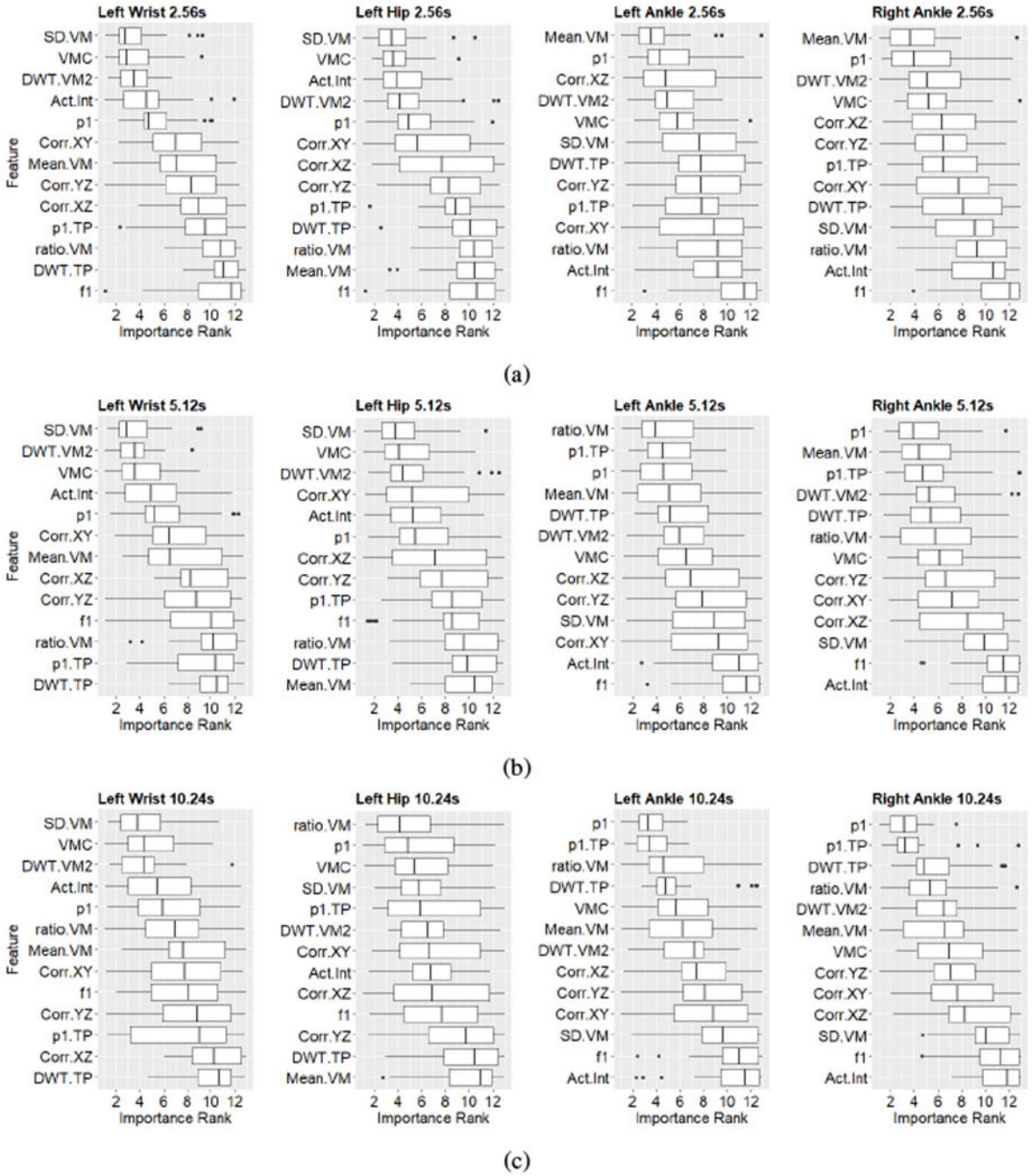

Fig. 7:

Variable importance rankings for the twelve scenarios for the subject-level classifiers. Variables are sorted from top to bottom by median importance rank.

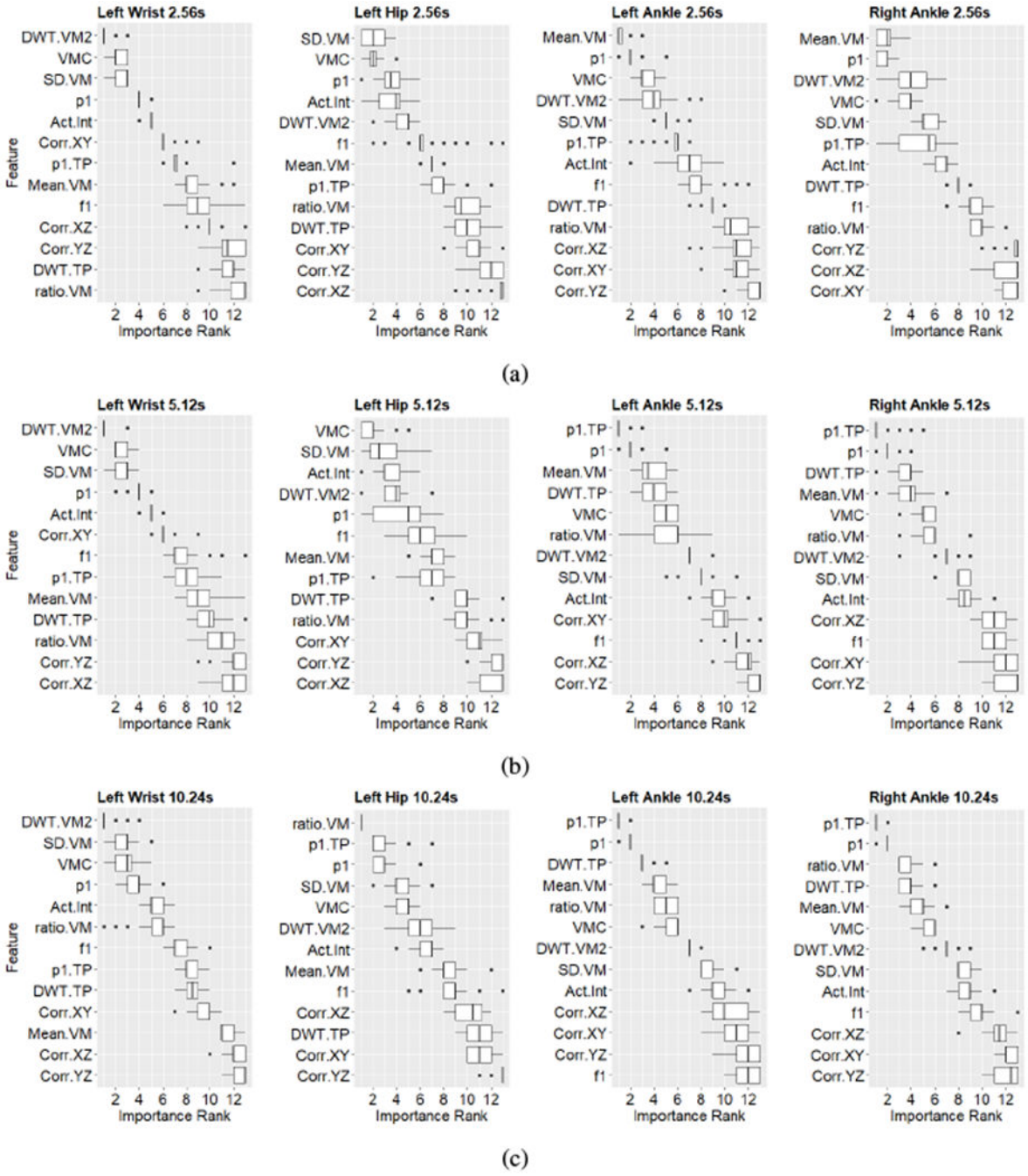

Fig. 8:

Variable importance rankings for the twelve scenarios for the population-level classifiers. Variables are sorted from top to bottom by median importance rank.

Table 1:

Study Demographics

Gender	N	Statistic	Mean	St. Dev.	Min	Max
Female	19	Age (y)	39.3	8.9	24.0	54.0
		Height(in)	65.8	3.7	58.0	73.0
		Weight(lbs)	143.0	32.1	100.0	212.0
		BMI (kg/m^2)	23.2	4.9	17.7	33.3
		Walk Time (mm:ss)	11:36	01:11	09:01	13:49
Male	13	Age (y)	38.6	9.5	23.0	52.0
		Height(in)	72.0	2.0	70.0	76.0
		Weight(lbs)	208.7	47.3	140.0	310.0
		BMI (kg/m^2)	28.2	5.5	20.1	39.8
		Walk Time (mm:ss)	11:31	00:58	09:47	13:01

Table 2:

Features extracted for walking classification

Feature	Description	Domain
$f1$	the dominant frequency between 1.2-4.0 Hz providing an estimate of the cadence (steps/second)	Frequency
ratio.VM	ratio of the partial area under the spectrum related to periodic movement to the complement	Frequency
$p1$	partial area under the power spectrum at $f1$	Frequency
$p1.TP$	ratio of $p1$ to the total area under the power spectrum between 0.3-12.5 Hz	Frequency
DWT.VM2	ratio of energy related to walking versus the total energy of the accelerometry signal	Frequency
DWT.TP	ratio of energy related to walking versus the total energy related to human movement	Frequency
VMC	vector magnitude count defined as the mean absolute deviation of the *VM*	Time
CORR.XY	correlation between the x- and y-axes of the accelerometry signal	Time
CORR.XZ	correlation between the x- and z-axes of the accelerometry signal	Time
CORR.YZ	correlation between the y- and z-axes of the accelerometry signal	Time
Act.Int	activity intensity defined as the average of the standard deviations for the x-, y-, and z-axes from the accelerometry signal	Time
Mean.VM	mean of the vector magnitude	Time
SD.VM	standard deviation of the vector magnitude	Time