Article

Employing Machine Learning to Estimate Hallmark Measures of Physical Activities from Wrist-worn Devices Across Age Groups

Mamoun Mardini [1, 2]*, Chen Bai [2], Amal A. Wanigatunga [3], Santiago Saldana [4], Ramon Casanova [4], Todd Manini [1]

[1] Department of Aging and Geriatric Research, College of Medicine, University of Florida, FL, USA; malmardini@ufl.edu (MM); tmanini@ufl.edu (TM)
[2] Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, FL, USA; chenbai@ufl.edu (CB)
[3] Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, MD, USA; awaniga1@jhu.edu
[4] Department of Biostatistics and Data Science, School of Medicine, Wake Forest University, NC, USA; ssaldana@wakehealth.edu(SS); casanova@wakehealth.edu(RC)

* Correspondence: malmardini@ufl.edu;

Abstract: Wrist-worn fitness trackers and smartwatches are proliferating with an incessant attention towards health tracking. Given the growing popularity of wrist-worn devices across all age groups, a rigorous evaluation for recognizing hallmark measures of physical activities and estimating energy expenditure is needed to compare their accuracy across the lifespan. The goal of the study was to build machine learning models to recognize physical activity type (sedentary, locomotion, and lifestyle) and intensity (low, light, and moderate), identify individual physical activities, and estimate energy expenditure. The primary aim of this study was to build and compare models for different age groups: young [20-50 years], middle (50-70 years], and old (70-89 years]. Participants (n = 253, 62% women, aged 20-89 years old) performed a battery of 33 daily activities in a standardized laboratory setting while wearing a portable metabolic unit to measure energy expenditure that was used to gauge metabolic intensity. Tri-axial accelerometer collected data at 80-100 Hz from the right wrist that was processed for 49 features. Results from random forests algorithm were quite accurate in recognizing physical activity type, the F1-Score range across age groups was: sedentary [0.955 – 0.973], locomotion [0.942 – 0.964], and lifestyle [0.913 – 0.949]. Recognizing physical activity intensity resulted in lower performance, the F1-Score range across age groups was: sedentary [0.919 – 0.947], light [0.813 – 0.828], and moderate [0.846 – 0.875]. The root mean square error range was [0.835 – 1.009] for the estimation of energy expenditure. The F1-Score range for recognizing individual physical activities was [0.263 – 0.784]. Performances were relatively similar and the accelerometer data features were ranked similarly between age groups. In conclusion, data features derived from wrist worn accelerometers lead to high-moderate accuracy estimating physical activity type, intensity and energy expenditure and are robust to potential age-differences.

Keywords: wrist; accelerometer; physical activity; energy expenditure; machine learning; random forest, age groups

1. Introduction

Regular and sufficient amounts of physical activity (PA) are significant in increasing health benefits and mitigating health risks. Globally, one out of four adults (almost 1.4 billion) do not meet the World

Health Organization (WHO) PA recommendations [1]. Mobility is an essential factor for independence and social life engagement. Those who lose mobility have higher risk of morbidity, disability, and mortality [2–5]. Recently, WHO has published the *Global action plan on physical activity 2018–2030* (GAPPA) to enhance PA with a target of 15% reduction in physical inactivity by year 2030 [6]. The most recent WHO guidelines on physical activity and sedentary behavior [7] suggest that adults (aged 18 and older) should do at least 150–300 minutes of moderate-intensity aerobic PA; or at least 75–150 minutes of vigorous intensity aerobic PA; or an equivalent combination of moderate- and vigorous-intensity activity throughout the week. Additionally, adults should replace their time spent being sedentary with PA.

To meet the WHO goals, accurate estimation of physical activity type, intensity and duration is required. The proliferation of fitness trackers and wearable accelerometers offer an excellent opportunity to achieving this goal. The literature contains many examples of machine learning algorithms including decision tree [8], random forests [8,9], and bag-of- words [10] processing and modeling accelerometer data. However, these models are often limited to a specific age group (e.g., adults 20-40 yrs old). The looming question here is whether known age differences in movement patterns influence the performance of the machine learning models. There is a paucity of research to examine the differences between models built to recognize PA type and intensity, recognize individual PA, and estimate energy expenditure (EE) across different age groups. Such knowledge will be useful in deriving age-specific models that improve prediction accuracy.

Historically, the adopted approach used to recognize PA type and intensity, and to estimate energy expenditure (EE) relied on data collected from the hip position in standardized laboratory settings. The advantage of the hip over other positions is the proximity to the body’s center of the mass, offering a convenient and accurate approach for capturing ambulatory activity [11]. However, the hip position is riddled with patient/participant compliance issues and inability to gather 24 hour data [12]. Alternatively, the wrist position has become popular for collecting accelerometer data due to a rise in smartwatches, convenience, ability to capture sleep quality (24 hours) and enhanced compliance in research studies [13–16]. Unfortunately, despite the popularity of wrist-worn accelerometers, there is a paucity of models that are deemed viable for accurately assessing PA [17,18]. The use of the wrist position to recognize PA type and intensity and estimate EE is challenging due to its potential limitation in quantifying and capturing large lower limb movements and other lifestyle activities. Therefore, models that can accurately recognize PA type and intensity and estimate energy expenditure from the wrist are greatly needed to meet the current demand.

This study utilizes a large amount of high-resolution raw accelerometer data collected from the wrist position coupled with metabolic intensity assessed in 253 adults aged 20-89 years. An aggregated set of relevant features were used as an input to machine learning models to recognize PA type and intensity, identify individual PA, and estimate EE. Machine learning models developed on specific age groups (young [20,50], middle (50-70], and old (70-89]) were then compared to test the hypothesis that model performance varies across age-group. Results are expected to help evaluate whether machine learning models used to represent wrist-worn accelerometer data need to be tailored to known age-differences in movement and behavior to optimize their accuracy.

2. Materials and Methods

2.1. Participants

Participants were community dwelling adults 20+ years old who were able to read and speak English language, were welling to undergo all testing procedures, and their weight was stable in the last three months (+/-5 lbs). Two-hundred and fifty-three (253) of the 264 participants who were enrolled were included in the analysis. Those excluded either had: missing of start/end time of activities (6 participants), insufficient length of activity or missing values (3 participants), and missing demographic information (2 participants). Institutional Review Board at the University of Florida approved all study procedures, and all participants provided written informed consents before the study.

2.2. Prescribed Activities and Visits

The ChoresXL study methods have been described previously by our group [19,20]. Briefly, participants performed a battery of 33 typical daily activities that were categorized into activity types and intensities calculated post-facto from metabolic unit data (supplemental Table S1). Tasks were chosen because they mimic daily chores activities, common among most Americans, and they are consistent with average time spent in the 2010 American Time Use Survey [21]. All tasks were performed in a standardized laboratory setting with scripted instructions for approximately 8-10 minutes to achieve a steady state energy expenditure. Participants performed all tasks at their own speed and were ordered from lowest to highest metabolic demand to reduce transfer of high metabolic effects of one task to another. To ease burden and exhaustion, participants performed all tasks over four visits. However, some did not complete all visits. Overall, 213 participants attended all 4 visits, 21 attended 3 visits, 7 attended only 2 visits, and 12 attended only 1 visit. In total, there were 941 data collection visits.

2.3. Instrumentation

Participants wore an ActiGraph GT3X-BT monitors on their right wrists (ActiGraph Inc, Pensacola, FL). The ActiGraph GT3X-BT monitor is a tri-axial lightweight accelerometer that records accelerations in units of gravity (1 g) in perpendicular, anterior-posterior, and medio-lateral axes. Accelerometers were programmed to collect data at 100 Hz sampling rate. Participants also wore a 2 Kg portable metabolic unit that estimated energy expenditure using principles of indirect calorimetry, Cosmed K5 (COSMED, Rome, Italy). Before data collection, the oxygen (O_2) and carbon dioxide (CO_2) sensors were calibrated using a gas mixture sample of 16.0% O_2 and 5.0% CO_2 and room air calibration. The turbine flow meter was calibrated using a 3.0-L syringe. A flexible facemask was positioned over the participant's mouth and nose and attached to the flow meter. Oxygen consumption ($VO_2 = mL.min^{-1}.kg^{-1}$) was measured breath-by-breath and were subsequently smoothed with a 30-sec running average window. Steady-state VO_2 for each task was manually calculated over approximately 2 minutes when there was evidence of a plateau, which indicates metabolic demand is matched to physical workload. Data were expressed as METs after dividing the VO_2 values by the traditional standard of 3.5 $mL.min^{-1}.kg^{-1}$ [22].

2.4. Problem Formulation

In this paper, we targeted four main tasks to measure the hallmark measures of PA: 1) recognize PA type (classification task) through splitting this task into three binary classification tasks: i) sedentary vs non-sedentary; ii) locomotion vs non-locomotion and iii) lifestyle vs non-lifestyle; 2) recognize PA intensity (classification task) through splitting this task into three binary classification tasks: i) low vs non-low; ii) light vs non-light and iii) moderate vs non-moderate; 3) recognize individual PA (classification task); and 4) estimate the energy expenditure while performing the scripted activities (regression task). We extracted consecutive non-overlapping 60-seconds windows from the raw accelerometer data. Previous studies used various window lengths, ranging from 0.1 seconds to 128 seconds [23–27]. A 60-seconds window was chosen as a compromise between having sufficient data for accurate feature extraction and balancing computational resources. In total, 49 time– and frequency–domain features, listed in Table 1, were extracted. During data processing, some cases with different collection frequencies were discovered (15 at 80 Hz and 100 at 30 Hz). However, no resampling was performed because the resolution was sufficient to extract features over a 60 second window.

Table 1. Description of features extracted from the raw data

	Feature	Description
Time	Mean of vector magnitude (mvm)	Sample mean of the VM in the window
	SD of vector magnitude (sdvm)	Standard deviation of VM
	Mean angle of acceleration relative to vertical on the device (mangle)	Sample mean of the angle between x axis and VM in the window
	SD of the angle of acceleration relative to vertical on the device (sdangle)	Sample standard deviation of the angles in the window
	Mean of acceleration (mean_x, mean_y and mean_z)	Sample mean of acceleration from x axis, y axis and z axis in the window
	SD of acceleration (sd_x, sd_y and sd_z)	Standard deviation of acceleration from x axis, y axis and z axis in the window
	Coefficient of variation of acceleration (cv_x, cv_y and cv_z)	Standard deviation of acceleration from x axis, y axis and z axis in the window divided by their mean, multiplied by 100
	Min of vector magnitude and acceleration (min_vm, min_x, min_y and min_z)	Min value of VM and acceleration from x axis, y axis and z axis in the window
	Max of vector magnitude and acceleratioin (max_vm, max_x, max_y and max_z)	Max value of VM and acceleration from x axis, y axis and z axis in the window
	25% quantile of vector magnitude and acceleratioin (lower_vm_25, lower_x_25, lower_y_25 and lower_z_25)	Lower 25% quantile of VM and acceleration from x axis, y axis and z axis in the window

	75% quantile of vector magnitude and acceleration (upper_vm_75, upper_x_75, upper_y_75 and upper_z_75)	Upper 75% quantile of VM and acceleration from x axis, y axis and z axis in the window
	Third moment of vector magnitude and acceleration (third_moment_vm, third_moment_x, third_moment_y and third_moment_z)	Third moment of VM and acceleration from x axis, y axis and z axis in the window, which are used to depict the shape of the signals
	Fourth moment of vector magnitude and acceleration (fourth_moment_vm, fourth_moment_x, fourth_moment_y and fourth_moment_z)	Fourth moment of VM and acceleration from x axis, y axis and z axis in the window, which are used to depict the shape of the signals
	Skewness	Skewness of the VM, acceleration from x axis, y axis, and z axis in the window
	Kurtosis	Kurtosis of the VM, acceleration from x axis, y axis and z axis in the window
	Coefficient of variation (CV)	Standard deviation of VM in the window divided by the mean, multiplied by 100
Frequency	Percentage of the power of the vm that is in 0.6-2.5 Hz (p625)	Sum of moduli corresponding to frequency in this range divided by sum of moduli of all frequencies
	Dominant frequency of vm (df)	Frequency corresponding to the largest modulus
	Fraction of power in vm at dominant frequency (fpdf)	Modulus of the dominant frequency/sum of moduli at each frequency

2.5. Model Training

Three main models were developed to estimate PA type recognition, PA intensity recognition, and individual PA recognition. The models were generated separately across three age groups: young [20-50 years], middle (50-70 years], and old (70-89 years]. For EE estimation, 247 participants provided valid data and were included. All the scripted activities (33 activities) were used in case of individual PA recognition, PA intensity recognition and EE estimation. However, for PA type recognition, some activities were removed (*strength exercise leg extension, strength exercise chest press, strength exercise leg curl, stretching yoga*); they did not fit sedentary, locomotion or lifestyle categories. For the PA type recognition, we built binary classification models for each type and age group; resulting in 12 models. Similarly, for the PA intensity recognition, we built binary classification models for each intensity and age group; resulting in 12 models. For individual PA recognition, we built one multi-class classification model (33 classes) for each age group; resulting in 4 models. For EE estimation, we built one regression model for each group; resulting in 4 models. In all tasks, all participants were randomly distributed into 5 folds. We used 5-fold nested cross validation (nested-CV), which has an inner CV loop nested in an outer CV loop. The inner loop is responsible for hyperparameter tuning (the process of searching for the optimal parameters of the model), while the outer loop is responsible for error estimation and generalization. We used random search for hyperparameter tuning (*number of trees, maximum number of features, maximum depth of each tree, and minimum number of samples per leaf*) , in which 10 sets of hyperparameters are set up and combined randomly for training the model. Then, the model with the highest F1-score was chosen. F1-score was used to compare across age groups because it protects against the imbalance across classes seen in PA type and intensity categories. There is no absolute criterion for a "good" value of F1 measure, but values above 0.80 generally indicate good performance. For continuous data from energy expenditure (METs), the root mean square error (RMSE) was used to evaluate performance.

3. Results

Table 2 shows participants' descriptive characteristics per age group: young [20-50 years], middle (50-70 years], and old (70-89 years]. Table 3 shows a slight performance reduction from younger to older age groups and from sedentary to more high variability lifestyle activities (F1-score range [0.913 – 0.973]). Results in Table 3 also show that METs RMSE decreased (improved) from young to middle to older age groups (RMSE range [0.835 - 1.009]).

Results for PA intensity show model performance were slightly higher for young and middle age groups compared to the old age group (F1-score range [0.813 – 0.947]). The performance of low intensity models across age groups outperformed the performance of the moderate, then light intensities.

Table 5 shows the performance of recognizing individual PA. It can be noticed that activities mainly involving wrist movements (washing dishes, computer work, cleaning windows) tend to perform better than others. However, there is no clear difference across age groups.

Figures 1-3 show the confusion matrices of recognizing PA type across age groups. The confusion increases as we move from sedentary to lifestyle PA type, which is consistent with the F1 scores shown in Table 3. Figures 4-6 show the confusion matrices of recognizing PA intensity across age groups. Similarly, the confusion of the models are consistent with the F1 scores shown in Table 4.

Figures 7-9 show the top 15 features that contributed the most in recognizing PA type across age groups. It can be noticed that the ranking of features is similar across age groups within each PA type. Figures 10-12 show the top 15 features that contributed the most in recognizing PA intensity across age groups. Similarly, it can be noticed that the ranking of features is similar across age groups within each PA intensity.

Table 2. Participants descriptive characteristics by age group

	Young	**Middle**	**Old**	**All**
Age range, years	[20-50]	(50-70]	(70-89]	[20-89]
Mean Age (SD), years	35.2 (10.7)	61.9 (5.6)	77.7 (5.1)	61.7 (17.7)

Mean BMI (SD), kg/m^2	26.1 (5.5)	26.9 (5.5)	27.7 (5.8)	27 (5.6)
Women %	60%	67%	58%	62%
Number of Hispanic	3	2	1	6
Total number	60	95	98	253

Table 3. Performance metrics of recognizing physical activity type and estimating energy expenditure. Each value is the mean and standard deviation of the 5-fold nested cross validation.

	Young	**Middle**	**Old**	**All**
Activity Type Recognition Performance (F1 Score)				
Sedentary	0.970 (0.004)	0.973 (0.004)	0.955 (0.005)	0.971 (0.002)
Locomotion	0.964 (0.009)	0.956 (0.004)	0.942 (0.005)	0.956 (0.004)
Lifestyle	0.949 (0.008)	0.940 (0.005)	0.913 (0.005)	0.938 (0.004)
Macro average (F1-score)*	0.961 (0.005)	0.956 (0.003)	0.937 (0.003)	0.955 (0.003)
Energy Expenditure Estimation Performance (RMSE)				
	1.009 (0.059)	0.904 (0.024)	0.835 (0.033)	0.898.048)

*The macro-average F1-score is the unweighted average of the F1-scores over all the classes

Table 4. Performance metrics of recognizing physical activity intensity. Each value is the mean and standard deviation of the 5-fold nested cross validation.

	Young	**Middle**	**Old**	**All**
Activity Intensity Recognition Performance (F1 Score)				
Low intensity	0.947 (0.014)	0.939 (0.005)	0.919 (0.012)	0.927 (0.005)
Light intensity	0.828 (0.008)	0.845 (0.013)	0.813 (0.010)	0.839 (0.004)
Moderate intensity	0.866 (0.012)	0.875 (0.012)	0.846 (0.011)	0.868 (0.005)
Macro average (F1-score)*	0.880 (0.008)	0.886 (0.010)	0.860 (0.009)	0.878 (0.004)

*The macro-average F1-score is the unweighted average of the F1-scores over all the classes

Table 5. Performance metrics of recognizing individual physical activities and estimating energy expenditure. Each value is the mean and standard deviation of the 5-fold nested cross validation.

	Young	**Middle**	**Old**	**All**
Individual Activities Recognition Performance (F1 Score)				
LEISURE WALK	0.510 (0.066)	0.488 (0.078)	0.401 (0.049)	0.474 (0.033)

RAPID WALK	0.631 (0.023)	0.527 (0.063)	0.463 (0.075)	0.557 (0.035)
LIGHT GARDENING	0.537 (0.090)	0.442 (0.058)	0.431 (0.035)	0.499 (0.041)
YARD WORK	0.349 (0.022)	0.383 (0.070)	0.351 (0.063)	0.405 (0.042)
PREPARE SERVE MEAL	0.415 (0.042)	0.379 (0.048)	0.380 (0.058)	0.428 (0.012)
DIGGING	0.664 (0.019)	0.634 (0.036)	0.605 (0.072)	0.661 (0.039)
STRAIGHTENING UP DUSTING	0.402 (0.075)	0.374 (0.053)	0.365 (0.059)	0.401 (0.028)
WASHING DISHES	0.724 (0.056)	0.661 (0.017)	0.583 (0.042)	0.671 (0.020)
UNLOADING STORING DISHES	0.668 (0.040)	0.627 (0.051)	0.557 (0.053)	0.637 (0.011)
WALKING AT RPE 1	0.316 (0.061)	0.456 (0.042)	0.300 (0.059)	0.395 (0.018)
PERSONAL CARE	0.644 (0.058)	0.662 (0.038)	0.492 (0.046)	0.622 (0.005)
DRESSING	0.445 (0.056)	0.384 (0.055)	0.263 (0.026)	0.386 (0.025)
WALKING AT RPE 5	0.415 (0.038)	0.397 (0.097)	0.348 (0.088)	0.393 (0.022)
SWEEPING	0.557 (0.061)	0.554 (0.084)	0.458 (0.053)	0.546 (0.021)
VACUUMING	0.592 (0.043)	0.550 (0.043)	0.493 (0.032)	0.551 (0.024)
STAIR DESCENT	0.643 (0.090)	0.610 (0.040)	0.576 (0.030)	0.613 (0.070)
STAIR ASCENT	0.422 (0.110)	0.486 (0.065)	0.437 (0.027)	0.482 (0.049)
TRASH REMOVAL	0.403 (0.054)	0.415 (0.066)	0.274 (0.057)	0.396 (0.028)
REPLACING SHEETS ON A BED	0.579 (0.069)	0.606 (0.059)	0.549 (0.016)	0.607 (0.029)
STRETCHING YOGA*	0.586 (0.041)	0.600 (0.029)	0.501 (0.062)	0.556 (0.031)
MOPPING	0.502 (0.076)	0.569 (0.041)	0.550 (0.077)	0.606 (0.033)
LIGHT HOME MAINTENANCE	0.447 (0.028)	0.466 (0.021)	0.378 (0.046)	0.468 (0.030)
COMPUTER WORK	0.764 (0.051)	0.784 (0.029)	0.750 (0.052)	0.782 (0.019)
HEAVY LIFTING	0.613 (0.060)	0.620 (0.015)	0.439 (0.059)	0.579 (0.019)
SHOPPING	0.447 (0.052)	0.488 (0.054)	0.442 (0.050)	0.486 (0.035)
IRONING	0.589 (0.028)	0.636 (0.027)	0.608 (0.057)	0.639 (0.035)
LAUNDRY WASHING	0.379 (0.035)	0.458 (0.036)	0.392 (0.021)	0.443 (0.030)
STRENGTH EXERCISE LEG CURL	0.560 (0.045)	0.616 (0.064)	0.652 (0.090)	0.652 (0.055)
STRENGTH EXERCISE CHEST PRESS	0.563 (0.051)	0.612 (0.085)	0.516 (0.066)	0.606 (0.037)
STRENGTH EXERCISE LEG EXTENSION	0.407 (0.152)	0.462 (0.067)	0.345 (0.063)	0.454 (0.034)
TV WATCHING	0.608 (0.054)	0.608 (0.036)	0.572 (0.079)	0.612 (0.038)
STANDING STILL	0.634 (0.120)	0.637 (0.079)	0.545 (0.097)	0.614 (0.036)
WASHING WINDOWS	0.740 (0.074)	0.693 (0.053)	0.722 (0.064)	0.729 (0.025)
Macro average (F1 score)	0.540 (0.025)	0.540 (0.025)	0.476 (0.011)	0.544 (0.015)

*The macro-average F1-score is the unweighted average of the F1-scores over all the classes

4. Discussion

The goal of the study was to build accurate machine learning models to recognizing the hallmark measures of physical activities and estimating energy expenditure across different age groups. We analyzed a large dataset of raw accelerometer data collected from the wrist position. We utilized the random forests algorithm, which is one of the most powerful algorithms in machine learning, to build models. Results showed that the machine learning models were quite accurate at recognizing physical activity type and intensity, and estimating energy expenditure. However, models performed less optimally when recognizing individual physical activities. Our hypothesis that increasing age would impact model performance was rejected as only slight differences were detected among age groups.

The results of the models built to recognize physical activity type showed high performance for all age groups as shown in Table 3. The model built on the young age group achieved the highest performance, followed by the middle, then old age groups for all activity types. Additionally, the highest performance was for sedentary, locomotion, then lifestyle activities for all age groups. Physical activity types seem to be more distinguishable and cause less confusion for younger ages as reflected on the confusion matrices shown in Figures 1-3. It is hard to interpret the drop in the performance from young to old age groups. One potential cause of this drop is the deviations from the standardized protocol that are more common in older adults. For example, there was a certain amount of variability in the *trash removal* activity among older adults compared to younger adults (older adults could not pull the trash bag quickly). This suggests that the ML models need to incorporate these compensations more accurately among older populations. Another reason is that older adults do not like the wrist device as tight as the younger adults. This can result in unintended artifactual movement that occurred more commonly among the older. Additional cause could be that the middle and old age groups include more participants' data than the young age group. Therefore, the models tend to generalize better and be less optimistic. On the other hand, the drop in the performance from sedentary to lifestyle activity types is intuitive. Lifestyle activities typically require more wrist involvement (i.e., ironing, trash removal) than other physical activity types. This means more variability in physical activities as we move from sedentary to lifestyle activities, which can increase the confusion in recognizing physical activity types as reflected in the confusion matrices shown in Figures 1-3.

The results of the models built to recognize physical activity intensity showed relatively high performance for all age groups, but lower than the performance of recognizing physical activity types as shown in Table 4. The highest performance was for the young and middle age groups alternatively, then old age group for all activity intensities. Additionally, the highest performance was for low, moderate, then light intensities for all age groups. As mentioned above, it is hard to interpret the drop in the performance from young to old age groups. Performance metrics and confusion for labeling physical activity intensities showed a consistent, although slight, reduction in older aged groups (see Table 4 and Figures 4-6). If this error was scaled to free-living conditions over a typical day (16 hours), older adults would be expected to have 2% (~19 minutes) more mislabeling of PA intensity compared to a younger group.

Models built to recognize individual physical activities showed lower performance than recognizing physical activity type. The highest F1-score was 0.784 in recognizing the computer work activity in the middle age group and the lowest was 0.263 for recognizing the dressing activity in the old age group. The overall deterioration in the recognition performance in individual activities compared to other recognition tasks is intuitive, due to the high number of classes and the data imbalance. Summing these activities into categories such as the physical activity types or physical activity intensities can help in enhancing the recognition performance metric as observed in Table 3 and Table 4. In general, there were no consistent differences among age groups.

The scaled impurity-based feature importance ranking generated from the random forest algorithm show how relevant these features are to the problem in hand and help in better understanding the model. We listed the top 15 features out of 49 features for both the physical activity type and intensity recognition tasks. By examining the feature importance for the physical activity types, there is a consistency in the ranking of these features across age groups within each one of the activity types. For example, variability in vector magnitude features such as *sdvm* and *cv_vm* were important in predicting sedentary physical activities, whereas wrist-specific features such as *wrist_sd_z* and *sd_angle* are more relevant for recognizing lifestyle activity types. The feature importance rankings for low intensity activities was similar to sedentary PA type, where the VM features such as *sdvm* and *cv_vm* were dominant. Feature rankings for predicting light and moderate intensities were similar with high importance for moment-based variables. Similarly, there is a consistency in the feature importance ranking across the age groups suggesting that the features are robust to potential movement difference with increasing age. Interestingly, the amplitude of the accelerometer axis (i.e. mean VM), which is commonly used to gauge intensity did not have a major role in model prediction. Being aware of the important features for the recognition problem in hand can help researchers continue improving model accuracy with less computational costs.

Comparing relevant literature results is an intricate endeavor because of the differences in the data collection environment and the variables that govern the study. There are numerous differences between studies that include: sample size, the demographic characteristics of participants, the number and diversity of the physical activities tested, type of accelerometer, body position, statistical and machine learning algorithms applied, the extracted statistical features, the window size, and the metrics measured to evaluate the overall performance. However, some important comparisons can be made. For example, Ellis et al. [28] built random forest models on data collected from the dominant wrist to predict physical activity type and estimate energy expenditure. The models were developed and tested on 40 (average age 35.8 years) participants. They obtained an average F1 score of 0.75 on 8 daily activities. Additionally, they obtained an RMSE value of 1.0 METs, which is similar to our young age group. Staudenmayer et al. [8] also used random forest to estimate energy expenditure and metabolic intensity of 19 physical activities from wrist accelerometer data. The models derived from a small young sample of 20 (24.1 years) estimated RMSE at 1.21 METs. When compared to others using machine learning approaches, the results from the current work are comparable within the young age group, but better in middle and old age groups.

Studies that examined the hallmark measures of physical activity have used publicly available data that contain activity labels, but not measures of metabolic intensity or energy expenditure (e.g. Opportunity (multiple body positions, 3 participants) [29], PAMAP2 (chest, arm and ankle positions, 9 participants) [30], UCI daily and sports dataset (hip position, 30 participants) [31], Skoda Mini Checkpoint (multiple body positions, 1 participant) [32], WISDM (hip position, 29 participants) [33], and Daphnet Freezing of Gait Dataset (legs and hip positions, 10 participants) [34]). They are also limited by a small number of participants, age-range being mostly < 40 years, a low number and diversity of activity types, and most importantly lacking sufficient data from the wrist position. Given these substantial differences, the models presented here show relatively higher performance than others. Additionally, the current model may generalize better due to the high diversity of activities, wide age-span, gender and racial diversity and the larger number of participants enrolled.

A limitation of the current study is that data were collected in controlled lab settings, which is appropriate and a first step in evaluating positional differences [35]. Collecting data in the free-living settings is more reflective of numerous transitions between activity types, but it is challenged by labeling the activity type. Another limitation is the consideration of window size, which was based on previous studies that extracted time- and frequency-domain features. This window size may not reflect the most appropriate size for all tasks and age groups. Additional simulation work should evaluate different window sizes for optimizing performance.

5. Conclusions

In this study, we tested the hypothesis that the machine learning model performance varies across age-groups for recognizing hallmark measures of physical activities and estimating energy expenditure. Overall results suggest data features derived from wrist worn accelerometers lead to high-to-moderate accuracy estimating physical activity type, intensity and energy expenditure in all age groups. In conclusion, machine learning models used to represent accelerometry data are robust to age differences and a generalizable approach might be sufficient to utilize in accelerometer-based devices (smartwatches and activity trackers).

Supplementary Materials:

Table S1: List of the performed physical activities, their type, and intensity

	Activity type			Intensity		
Activity	**Sedentary**	**Locomotion**	**Life-style**	**Low**	**light**	**Moderate**
LEISURE WALK	No	Yes	No	No	No	Yes
RAPID WALK	No	Yes	No	No	No	Yes
LIGHT GARDENING	No	No	Yes	No	No	Yes
YARD WORK	No	No	Yes	No	No	Yes
PREPARE SERVE MEAL	No	No	Yes	No	Yes	No
DIGGING	No	No	Yes	No	No	Yes
STRAIGHTENING UP	No	No	Yes	No	No	Yes

DUSTING						
WASHING DISHES	No	No	Yes	No	Yes	No
UNLOADING STORING DISHES	No	No	Yes	No	Yes	No
WALKING AT RPE 1	No	Yes	No	No	No	Yes
PERSONAL CARE	No	No	Yes	No	Yes	No
DRESSING	No	No	Yes	No	Yes	No
WALKING AT RPE 5	No	Yes	No	No	No	Yes
SWEEPING	No	No	Yes	No	No	Yes
VACUUMING	No	No	Yes	No	No	Yes
STAIR DESCENT	No	Yes	No	No	No	Yes
STAIR ASCENT	No	Yes	No	No	No	Yes
TRASH REMOVAL	No	No	Yes	No	No	Yes
REPLACING SHEETS ON A BED	No	No	Yes	No	No	Yes
STRETCHING YOGA*	No	No	No	No	Yes	No
MOPPING	No	No	Yes	No	No	Yes
LIGHT HOME MAINTENANCE	No	No	Yes	No	No	Yes
COMPUTER WORK	Yes	No	No	Yes	No	No
HEAVY LIFTING	No	No	Yes	No	No	Yes
SHOPPING	No	No	Yes	No	Yes	No
IRONING	No	No	Yes	No	Yes	No
LAUNDRY WASHING	No	No	Yes	No	Yes	No
STRENGTH EXERCISE LEG CURL*	No	No	No	No	Yes	No
STRENGTH EXERCISE CHEST PRESS*	No	No	No	No	Yes	No
STRENGTH EXERCISE LEG EXTENSION*	No	No	No	No	Yes	No
TV WATCHING	Yes	No	No	Yes	No	No
STANDING STILL	Yes	No	No	Yes	No	No
WASHING WINDOWS	No	No	Yes	No	No	Yes

A total of 29 activities were considered for PA type recognition, 33 for individual PA recognition, PA intensity recognition, and EE estimation.

* Only considered for energy expenditure estimation, PA intensity recognition, and individual PA recognition.

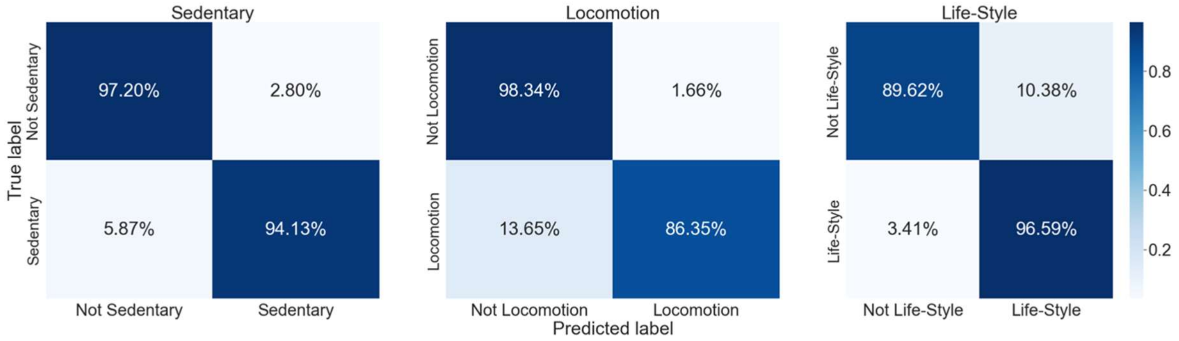

Figure 1. Confusion matrix of recognizing physical activity type for young age group

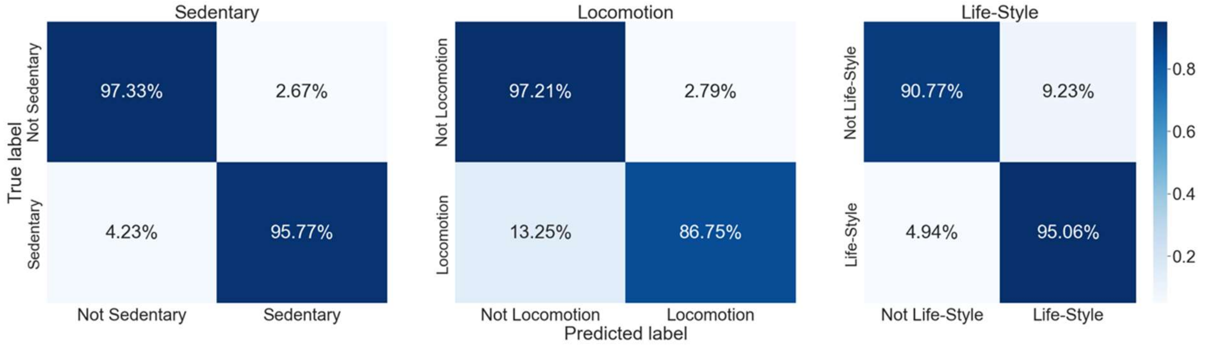

Figure 2. Confusion matrix of recognizing physical activity type for middle age group

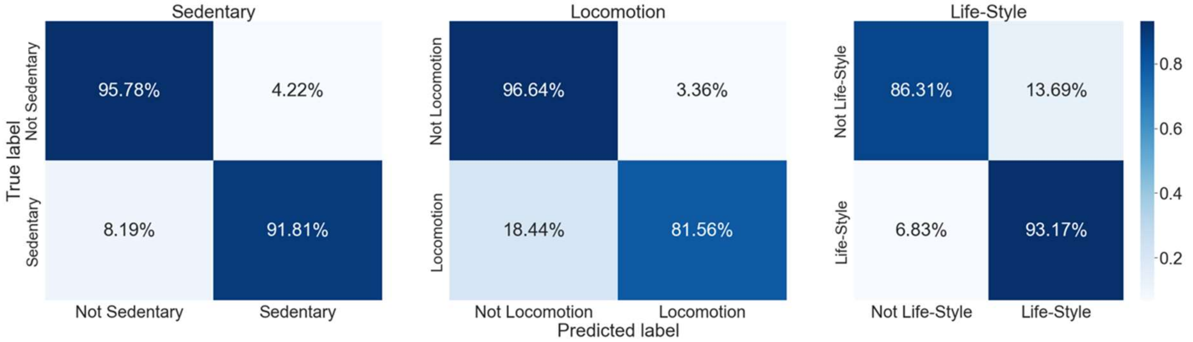

Figure 3. Confusion matrix of recognizing physical activity type for old age group

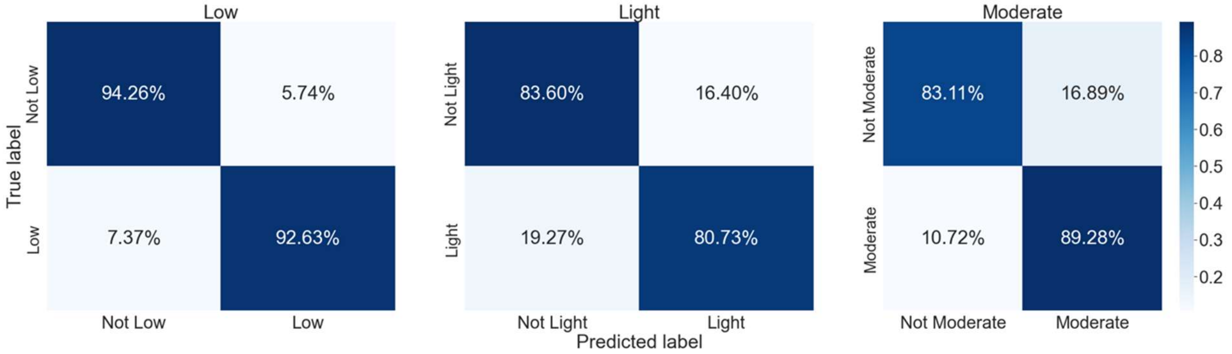

Figure 4. Confusion matrix of recognizing physical activity intensity for young age group

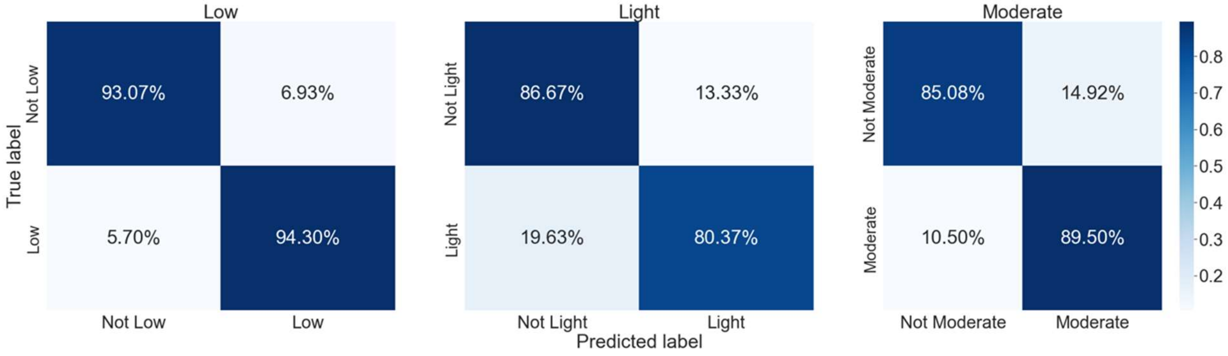

Figure 5. Confusion matrix of recognizing physical activity intensity for middle age group

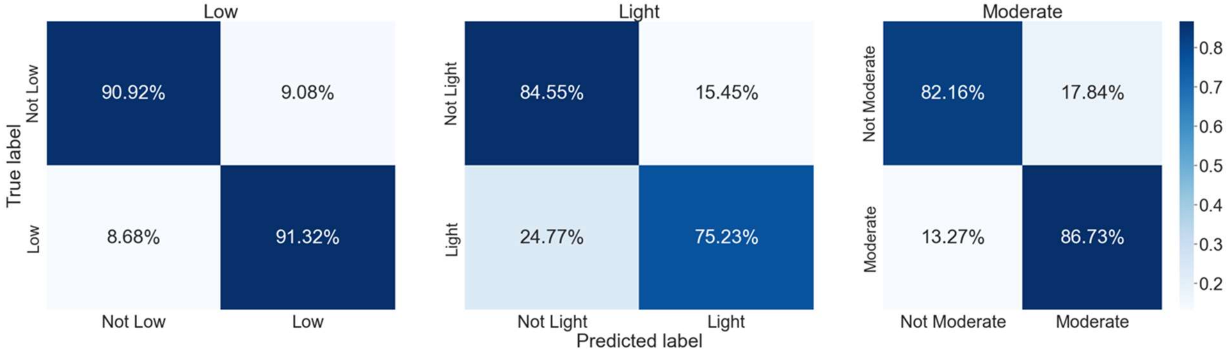

Figure 6. Confusion matrix of recognizing physical activity intensity for old age group

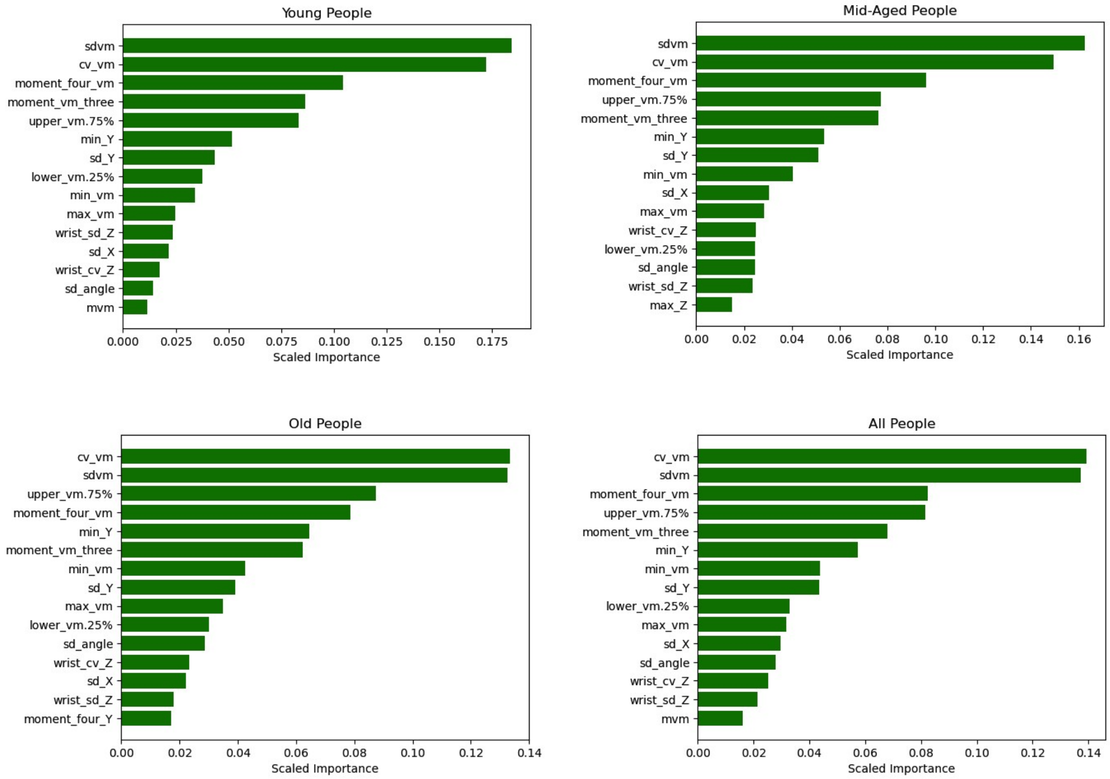

Figure 7. Feature importance for recognizing sedentary activities across age groups

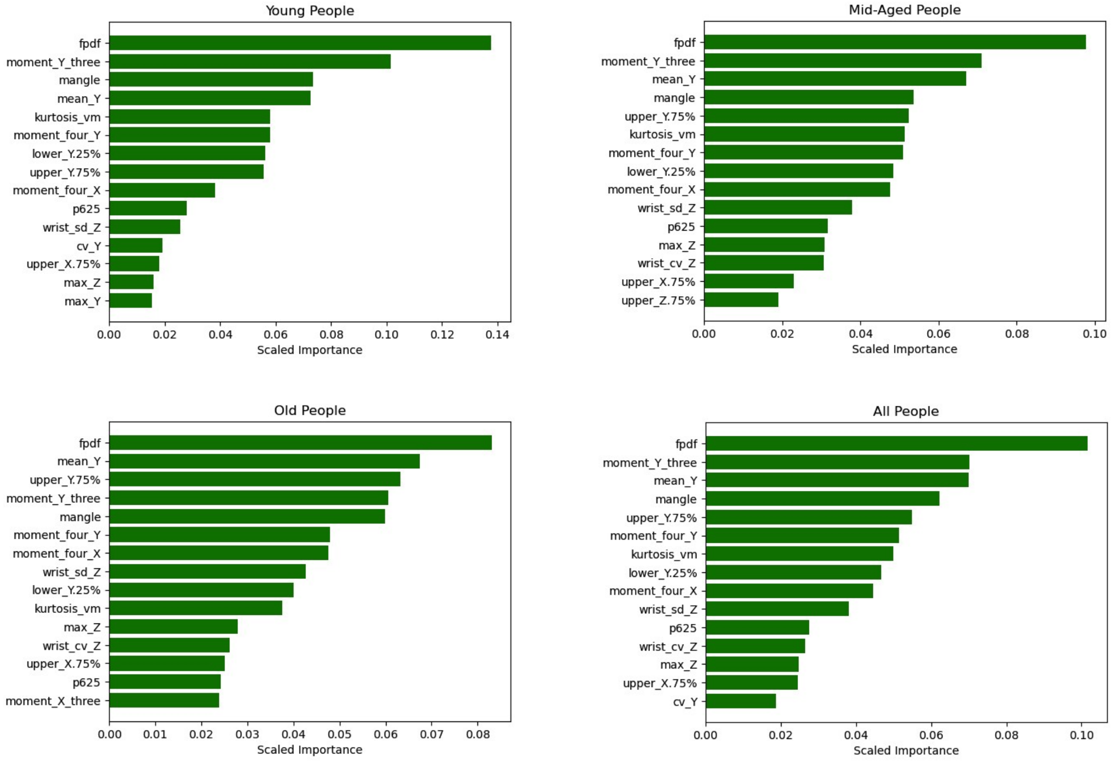

Figure 8. Feature importance for recognizing locomotion activities across age groups

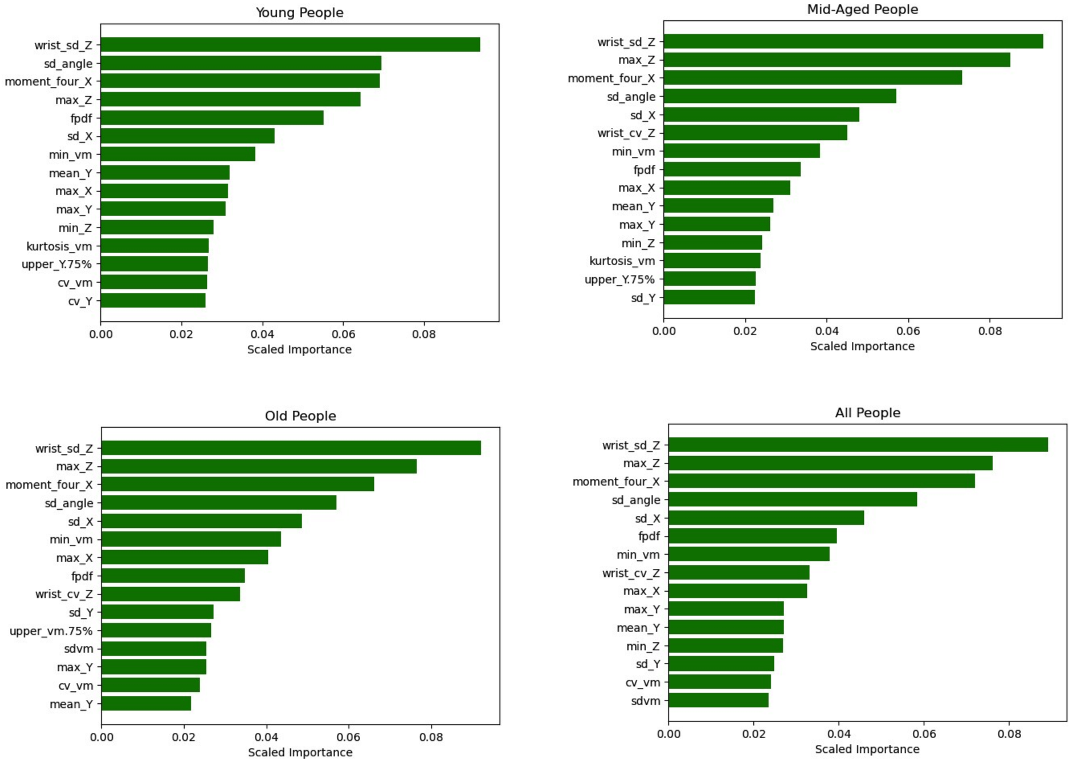

Figure 9. Feature importance for recognizing lifestyle activities across age groups

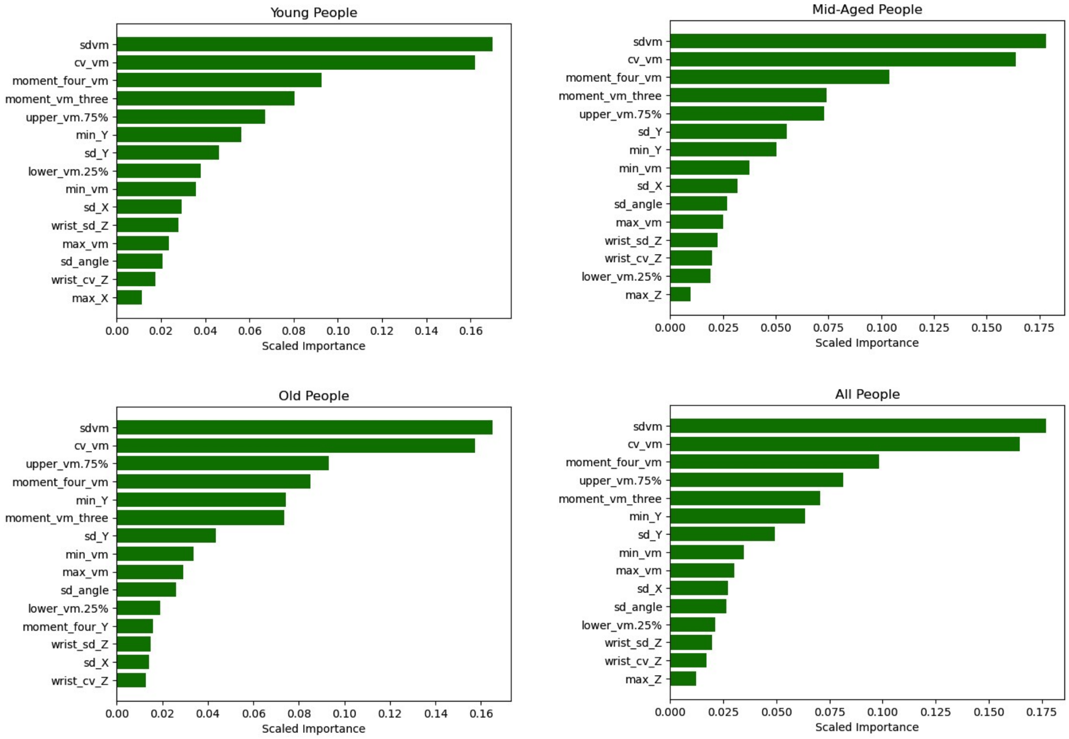

Figure 10. Feature importance for recognizing low intensity across age groups

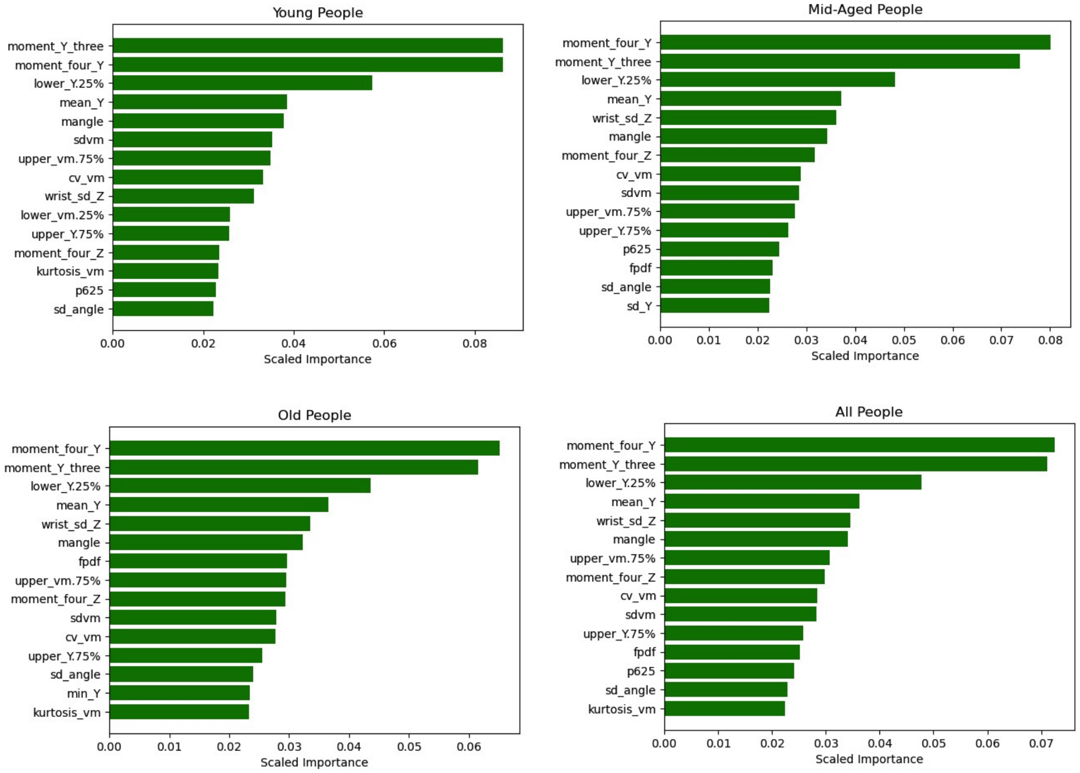

Figure 11. Feature importance for recognizing light intensity across age groups

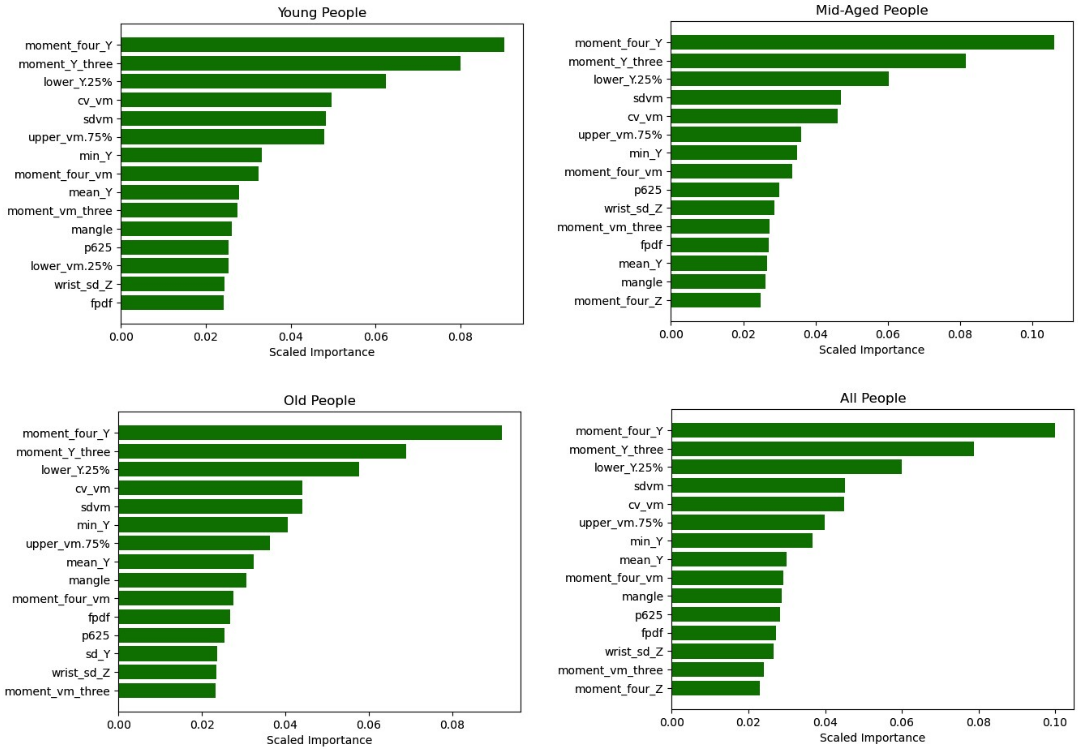

Figure 12. Feature importance for recognizing moderate intensity across age groups

Author Contributions: Conceptualization, M.M. and T.M.; methodology, M.M. and T.M.; software, S.S. and C.B.; validation, M.M. and T.M.; formal analysis, C.B. and S.S.; investigation, M.M. and T.M.; resources, T.M.; data curation, M.M., C.B., T.M. and S.S.; writing—original draft preparation, M.M. and T.M.; writing—review and editing, All; visualization, C.B.; supervision, T.M. and R.C.; project administration, T.M.; funding acquisition, T.M. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Institutes of Health (NIH)/National Institute on Aging (NIA) (R01AG042525), and partially supported by the Claude D. Pepper Older Americans Independence Centers at the University of Florida (P30AG028740).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] R. Guthold, G.A. Stevens, L.M. Riley, F.C. Bull, Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants, Lancet Glob. Heal. 6 (2018) e1077–e1086. doi:10.1016/S2214-109X(18)30357-7.

[2] L.G. Branch, A.M. Jette, A prospective study of long-term care institutionalization among the aged., Am. J. Public Health. 72 (1982) 1373–1379. doi:10.2105/AJPH.72.12.1373.

[3] M.C. Corti, J.M. Guralnik, M.E. Salive, L. Ferrucci, M. Pahor, R.B. Wallace, C.H. Hennekens, Serum iron level, coronary artery disease, and all-cause mortality in older men and women, Am. J. Cardiol. 79 (1997) 120–127. doi:10.1016/S0002-9149(96)00697-2.

[4] S.R. Khokhar, Y. Stern, K. Bell, K. Anderson, E. Noe, R. Mayeux, S.M. Albert, Persistent Mobility Deficit in the Absence of Deficits in Activities of Daily Living: A Risk Factor for Mortality, J. Am. Geriatr. Soc. 49 (2001) 1539–1543. doi:10.1046/j.1532-5415.2001.4911251.x.

[5] A.B. Newman, E.M. Simonsick, B.L. Naydeck, R.M. Boudreau, S.B. Kritchevsky, M.C. Nevitt, M. Pahor, S. Satterfield, J.S. Brach, S.A. Studenski, T.B. Harris, Association of long-distance corridor walk performance with mortality, cardiovascular disease, mobility limitation, and disability, J. Am. Med. Assoc. 295 (2006) 2018–2026. doi:10.1001/jama.295.17.2018.

[6] NCDs | WHO launches ACTIVE: a toolkit for countries to increase physical activity and reduce noncommunicable diseases, WHO. (2018). http://www.who.int/ncds/prevention/physical-activity/active-toolkit/en/ (accessed December 9, 2020).

[7] WHO guidelines on physical activity and sedentary behaviour, (n.d.). https://www.who.int/publications/i/item/9789240015128 (accessed January 15, 2021).

[8] J. Staudenmayer, S. He, A. Hickey, J. Sasaki, P. Freedson, Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements, J. Appl. Physiol. 119 (2015) 396–403. doi:10.1152/japplphysiol.00026.2015.

[9] K. Ellis, S. Godbole, S. Marshall, G. Lanckriet, J. Staudenmayer, J. Kerr, Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms, Front. Public Heal. 2 (2014). doi:10.3389/fpubh.2014.00036.

[10] M. Kheirkhahan, S. Mehta, M. Nath, A.A. Wanigatunga, D.B. Corbett, T.M. Manini, S. Ranka, A bag-of-words approach for assessing activities of daily living using wrist accelerometer data, in:

Proc. - 2017 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2017, Institute of Electrical and Electronics Engineers Inc., 2017: pp. 678–685. doi:10.1109/BIBM.2017.8217735.

[11] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, Y. Amirat, Physical Human Activity Recognition Using Wearable Sensors, Sensors. 15 (2015) 31314–31338. doi:10.3390/s151229858.

[12] K.Y. Troiano, Richard P, McClain, James J, Brychta, Robert J, Chen, Evolution of accelerometer methods for physical activity research, Br. J. Sports Med. 48 (2014) 1019–1023. doi:10.1136/bjsports-2014-093546.

[13] J.H. Migueles, C. Cadenas-Sanchez, U. Ekelund, C. Delisle Nyström, J. Mora-Gonzalez, M. Löf, I. Labayen, J.R. Ruiz, F.B. Ortega, Accelerometer Data Collection and Processing Criteria to Assess Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations, Sport. Med. 47 (2017) 1821–1845. doi:10.1007/s40279-017-0716-0.

[14] J. Kerr, C.R. Marinac, K. Ellis, S. Godbole, A. Hipp, K. Glanz, J. Mitchell, F. Laden, P. James, D. Berrigan, Comparison of Accelerometry Methods for Estimating Physical Activity, Med. Sci. Sports Exerc. 49 (2017) 617–624. doi:10.1249/MSS.0000000000001124.

[15] Worldwide Wearables Market Forecast to Maintain Double-Digit Growth in 2020 and Through 2024, According to IDC, (n.d.). https://www.idc.com/getdoc.jsp?containerId=prUS46885820 (accessed December 9, 2020).

[16] K.M. Full, J. Kerr, M.A. Grandner, A. Malhotra, K. Moran, S. Godoble, L. Natarajan, X. Soler, Validation of a physical activity accelerometer device worn on the hip and wrist against polysomnography, Sleep Heal. 4 (2018) 209–216. doi:10.1016/j.sleh.2017.12.007.

[17] H. Kinnunen, K. Häkkinen, M. Schumann, L. Karavirta, K.R. Westerterp, H. Kyröläinen, Training-induced changes in daily energy expenditure: Methodological evaluation using wrist-worn accelerometer, heart rate monitor, and doubly labeled water technique, PLoS One. 14 (2019) e0219563. doi:10.1371/journal.pone.0219563.

[18] O. Driscoll, How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies, Br J Sport. Med. 54 (2020) 332–340. doi:10.1136/bjsports-2018-099643.

[19] D.B. Corbett, A.A. Wanigatunga, V. Valiani, E.M. Handberg, T.W. Buford, B. Brumback, R. Casanova, C.M. Janelle, T.M. Manini, Metabolic costs of daily activity in older adults (Chores XL) study: Design and methods, Contemp. Clin. Trials Commun. 6 (2017) 1–8. doi:10.1016/j.conctc.2017.02.003.

[20] J.D. Knaggs, K.A. Larkin, T.M. Manini, Metabolic cost of daily activities and effect of mobility impairment in older adults, J. Am. Geriatr. Soc. 59 (2011) 2118–2123. doi:10.1111/j.1532-5415.2011.03655.x.

[21] American Time Use Survey Home Page, (n.d.). https://www.bls.gov/tus/ (accessed December 9, 2020).

[22] M. Jetté, K. Sidney, G. Blümchen, Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity, Clin. Cardiol. 13 (1990) 555–565. doi:10.1002/clc.4960130809.

[23] A. Krause, A. Smailagic, D.P. Siewiorek, J. Farringdon, Unsupervised, dynamic identification of

physiological and activity context in wearable computing, in: Proc. - Int. Symp. Wearable Comput. ISWC, IEEE Computer Society, 2003: pp. 88–97. doi:10.1109/iswc.2003.1241398.

[24] A. Mannini, S.S. Intille, M. Rosenberger, A.M. Sabatini, W. Haskell, Activity recognition using a single accelerometer placed at the wrist or ankle, Med. Sci. Sports Exerc. 45 (2013) 2193–2203. doi:10.1249/MSS.0b013e31829736d6.

[25] S. Pirttikangas, K. Fujinami, T. Nakajima, Feature selection and activity recognition from wearable sensors, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), Springer Verlag, 2006: pp. 516–527. doi:10.1007/11890348_39.

[26] M. Stikic, T. Huỳnh, K. Van Laerhoven, B. Schiele, ADL recognition based on the combination of RFID and aeeelerometer sensing, in: Proc. 2nd Int. Conf. Pervasive Comput. Technol. Healthc. 2008, PervasiveHealth, 2008: pp. 258–263. doi:10.1109/PCTHEALTH.2008.4571084.

[27] T. Huynh, B. Schiele, Analyzing features for activity recognition, in: ACM Int. Conf. Proceeding Ser., ACM Press, New York, New York, USA, 2005: pp. 159–164. doi:10.1145/1107548.1107591.

[28] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, S. Marshall, A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers, Physiol. Meas. 35 (2014) 2191–2203. doi:10.1088/0967-3334/35/11/2191.

[29] R. Chavarriaga, H. Sagha, A. Calatroni, S.T. Digumarti, G. Tröster, J.D.R. Millán, D. Roggen, The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition, Pattern Recognit. Lett. 34 (2013) 2033–2042. doi:10.1016/j.patrec.2012.12.014.

[30] A. Reiss, D. Stricker, Introducing a new benchmarked dataset for activity monitoring, in: Proc. - Int. Symp. Wearable Comput. ISWC, 2012: pp. 108–109. doi:10.1109/ISWC.2012.13.

[31] UCI daily and sports dataset, (n.d.). https://archive.ics.uci.edu/ml/index.php (accessed January 13, 2021).

[32] wiki:dataset [Human Activity/Context Recognition Datasets], (n.d.). http://har-dataset.org/doku.php?id=wiki:dataset (accessed January 13, 2021).

[33] WISDM Lab: Dataset, (n.d.). https://www.cis.fordham.edu/wisdm/dataset.php (accessed January 13, 2021).

[34] Daphnet Freezing of Gait Data Set, (n.d.). https://archive.ics.uci.edu/ml/index.php (accessed January 13, 2021).

[35] S.K. Keadle, K.A. Lyden, S.J. Strath, J.W. Staudenmayer, P.S. Freedson, A Framework to Evaluate Devices That Assess Physical Behavior, Exerc. Sport Sci. Rev. 47 (2019) 206–214. doi:10.1249/JES.0000000000000206.