Article

Age Differences in Estimating Physical Activity by Wrist Accelerometry Using Machine Learning

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

1 Department of Aging and Geriatric Research, College of Medicine, University of Florida, Gainesville, FL 32610, USA; tmanini@ufl.edu
2 Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL 32610, USA; chenbai@ufl.edu
3 Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA; awaniga1@jhu.edu
4 Department of Biostatistics and Data Science, School of Medicine, Wake Forest University, Winston-Salem, NC 27101, USA; ssaldana@wakehealth.edu (S.S.); casanova@wakehealth.edu (R.C.)
* Correspondence: malmardini@ufl.edu

Abstract: Accelerometer-based fitness trackers and smartwatches are proliferating with incessant attention towards health tracking. Despite their growing popularity, accurately measuring hallmark measures of physical activities has yet to be accomplished in adults of all ages. In this work, we evaluated the performance of four machine learning models: decision tree, random forest, extreme gradient boosting (XGBoost) and least absolute shrinkage and selection operator (LASSO), to estimate the hallmark measures of physical activities in young (20–50 years), middle-aged (50–70 years], and older adults (70–89 years]. Our models were built to recognize physical activity types, recognize physical activity intensities, estimate energy expenditure (EE) and recognize individual physical activities using wrist-worn tri-axial accelerometer data (33 activities per participant) from a large sample of participants (n = 253, 62% women, aged 20–89 years old). 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. F1-Scores derived from XGBoost's models were high for sedentary (0.955–0.973), locomotion (0.942–0.964) and lifestyle (0.913–0.949) activity types with no apparent difference across age groups. Low (0.919–0.947), light (0.813–0.828) and moderate (0.846–0.875) physical activity intensities were also recognized accurately. The root mean square error range for EE was approximately 1 equivalent of resting EE [0.835–1.009 METs]. Generally, random forest and XGBoost models outperformed other models. In conclusion, machine learning models to label physical activity types, activity intensity and energy expenditure are accurate and there are minimal differences in their performance across young, middle-aged and older adults.

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 the 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 min of moderate-intensity aerobic PA or at least 75–150 min 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 are required. The proliferation of fitness trackers and wearable accelerometers offer an excellent opportunity to achieve this goal. The literature contains many examples of machine learning algorithms used for the processing and modeling of the accelerometer data including decision tree [8], random forest [8,9], bag-of-words [10], neural network [11] and others [12–19]. However, these models are often limited to a specific age group (e.g., adults 20–40 years 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 [20]. However, the hip position is riddled with patient/participant compliance issues and inability to gather 24 h data [21]. Alternatively, the wrist position has become popular for collecting accelerometer data due to a rise in smartwatches, convenience, ability to capture sleep quality (24 h) and enhanced compliance in research studies [22–25]. Unfortunately, despite the popularity of wrist-worn accelerometers, there is a paucity of models that are deemed viable for accurately assessing PA [26,27]. 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 was 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, as shown in Figure 1. 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.
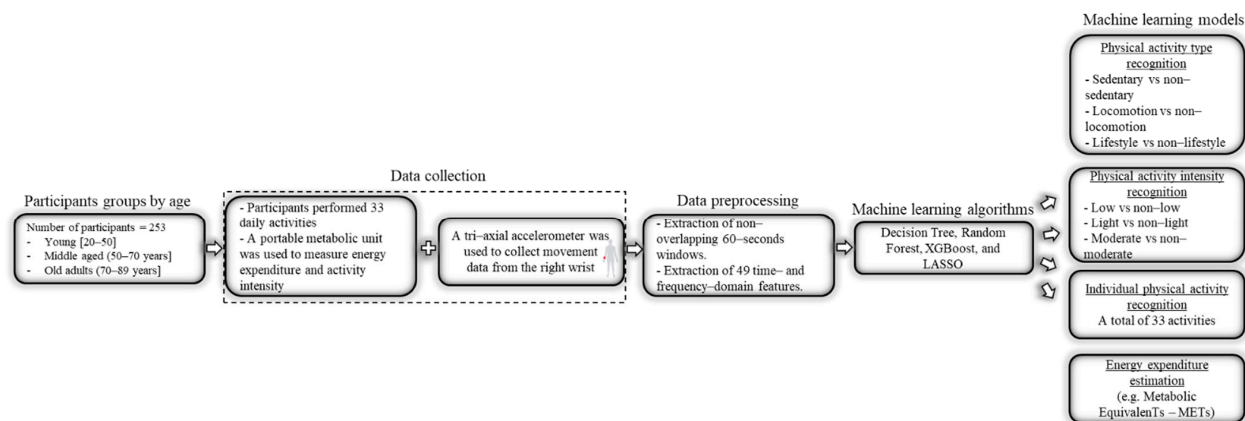

Figure 1. A block diagram showing the steps followed to collect and process the data.

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 willing 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 start/end time of activities (6 participants), insufficient length of activity or missing values (3 participants) and missing demographic information (2 participants). The 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 [28,29]. 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 [30]. All tasks were performed in a standardized laboratory setting with scripted instructions for approximately 8–10 min 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 ActiGraph GT3X-BT monitors on their right wrists (ActiGraph Inc, Pensacola, FL, USA). 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 was subsequently smoothed with a 30-s running average window. Steady-state VO_2 for each task was manually calculated over approximately 2 min 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} [31].

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-s windows from the raw accelerometer data. Previous studies used various window lengths, ranging from 0.1 s to 128 s [32–35]. A 60-s 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. Although there is inconsistency among researchers about the aggregation of relevant features, in this work, we combined features from previous studies [8,36–40]. 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-s 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 acceleration (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 acceleration (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

Four main models were developed for PA type recognition, PA intensity recognition, EE estimation 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 the 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. We utilized four algorithms to build our machine learning models: decision tree [40], random forest [40], extreme gradient boosting (XGBoost) [41] and least absolutes and selection operator (LASSO) [42]. Our selection of these algorithms was based on the fact that all the models provide better interpretability and include feature selection as part of the model building process. For the PA type recognition, we built binary classification models for each type and age group, resulting in 48 models. Similarly, for the PA intensity recognition, we built binary classification models for each intensity and age group, resulting in 48 models. For individual PA recognition, we built one multi-class classification model (33 classes) for each age group using the best performing model (XGBoost), resulting in 4 models. For EE estimation, we built one regression model for each group, resulting in 16 models. All the utilized machine learning algorithms are naturally resistant to insignificant predictors. They intrinsically perform feature selection to enhance the predictability of the models [43]. In all tasks, all participants were randomly distributed into 5 folds. In order to solve the data imbalance, the models were set to automatically adjust weights inversely proportional to the frequencies of the classes in the input data. We used 5-fold nested cross validation (nested-CV), which uses a series of train, validation and test set splits. The nested CV consists of 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. Initially, the data is split into outer training and testing datasets (outer loop). Then, the outer training dataset is further split into inner training and testing datasets (inner loop). Validation and hyperparameter tuning happen in the inner datasets, then the performance is reported on the outer testing datasets. This process was repeated 5 times. Then, the model with the highest performance was chosen. In this approach, the model selection becomes an integral part of the model fitting process, which results in preventing bias in performance evaluation [44–46]. We used four metrics to report the models' performance: F1-Score = 2 × (precision × recall)/(precision + recall), area under the curve (AUC) = area under true positive rate vs. false positive rate, balanced accuracy = (sensitivity + specificity)/2 and accuracy. The F1-Score measures the harmonic mean of precision and recall. The 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 the continuous data from energy expenditure (METs), the root mean square error (RMSE) was used to evaluate performance.

2.6. Brief Overview of the Utilized Machine Learning Algorithms

LASSO (least absolute shrinkage and selection operator) is a statistical and machine learning regression algorithm used for feature selection and regularization to enhance the performance of the model. It is a type of linear regression that utilizes the concept of shrinkage, in which the data points are shrunk towards a central point [47]. We utilized the *Glmnet* package [47], in which the LASSO linear and logistic regression are implemented.

Decision tree learning is one of the most common machine learning algorithms due to its simplicity and interpretability. The tree is a graphical representation of decisions. It consists of leaves representing the class labels (e.g., sedentary or non-sedentary), and branches representing conjunctions of features (e.g., time- and frequency-domain features)

that lead to those class labels. The tree is built by splitting the source dataset into subsets. Each subset is used to select the feature that best split the data equally. Decision tree learning can be used for building classification trees (e.g., PA type and intensity recognition) or regression trees (e.g., EE estimation) [48].

Random forest is an ensemble learning algorithm based on the concept of bagging (or bootstrap aggregation), in which predictions from multiple decision trees are combined through a majority voting mechanism. In random forest, however, only a subset of features are selected randomly to build a forest of decision trees [40].

XGBoost (extreme gradient boosting) is also an ensemble learning algorithm based on the gradient boosting framework, in which models are built sequentially to boost (increase) the performance of the previous models by utilizing the gradient descent algorithm to minimize errors. However, XGBoost offers better hardware and software optimization mechanisms, prevents overfitting by penalizing complex models and handles sparse patterns and missing data efficiently [41].

3. Results

Table 2 shows participants' descriptive characteristics per age group: young (20–50 years), middle (50–70 years) and old (70–89 years). There are no noticeable differences between the different age groups with respect to BMI, women percentage and number of Hispanic. Figure 2 shows a slight performance reduction from younger to older age groups and from sedentary to more high variability lifestyle activities. The performance of the machine learning models were similar in recognizing sedentary PA type and varied in recognizing locomotion and lifestyle PA types. Generally, XGBoost and random forest models outperformed other models. However, the XGBoost models were slightly better than the random forest models in most of the tasks, except for recognizing sedentary PA type across all age groups. Figures S1–S3 show other performance metrics including AUC, balanced accuracy and accuracy.
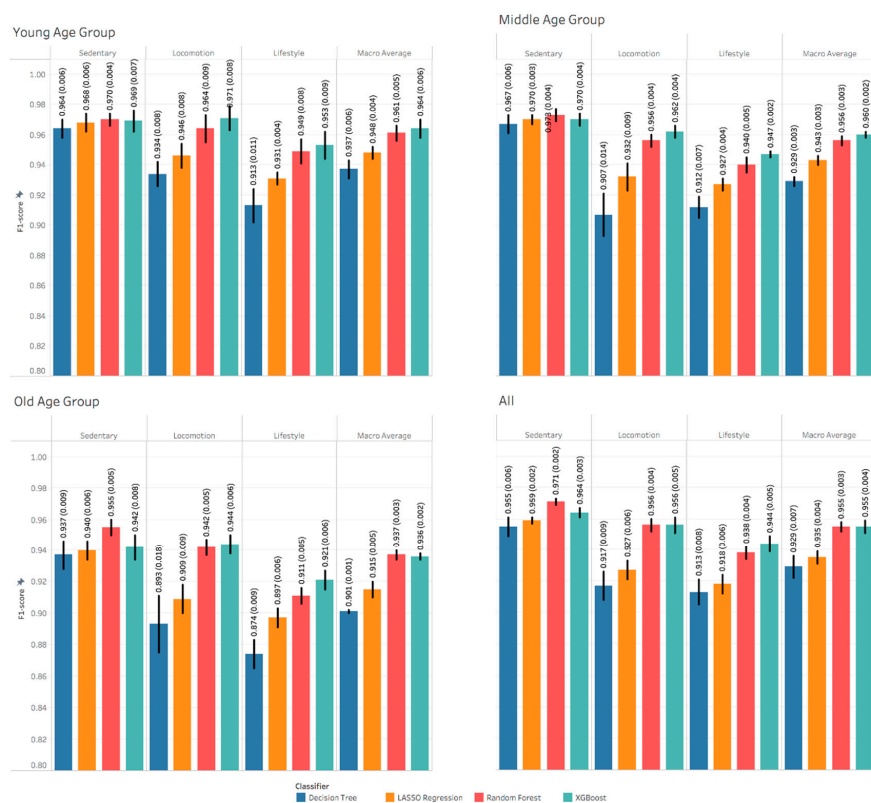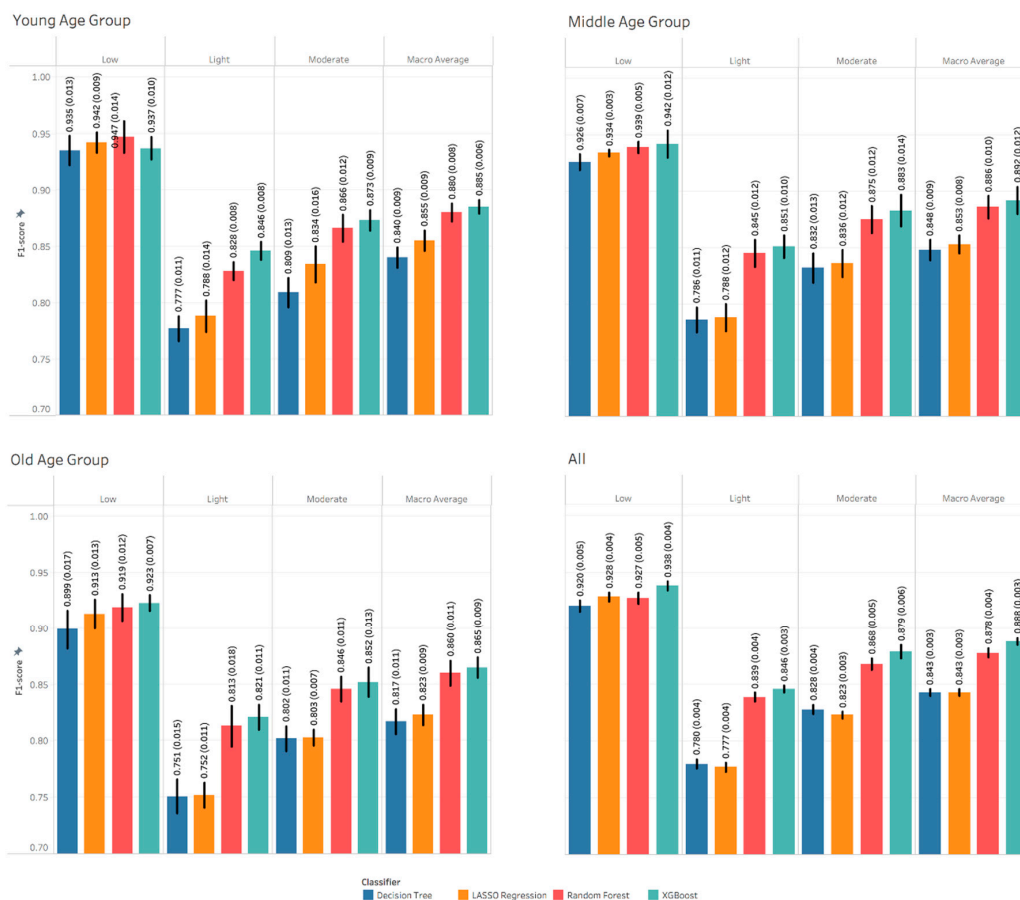

Figure 2. The F1-Scores of recognizing physical activity type. Each value is the mean and standard deviation of the 5-fold nested cross validation.

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

Results for PA intensity show that the models' performance was slightly higher for young and middle age groups compared to the old age group, as shown in Figure 3. The performance of low intensity models across age groups outperformed the performance of the moderate, then light intensities. The performance of the machine learning models were close in recognizing low PA intensity and varied in recognizing light and moderate PA intensities. Generally, XGBoost and random forest models outperformed other models. However, the XGBoost models were slightly better than the random forest models in most of the tasks, except for recognizing low PA intensity in the young age group. Figures S4–S6 show other performance metrics including AUC, balanced accuracy and accuracy.

Figure 3. Performance metrics of recognizing physical activity intensity. Each value is the mean and standard deviation of the 5-fold nested cross validation.

Figure 4 shows that METs RMSE decreased (improved) from young to middle to older age groups. The performance of the machine learning models was close for the young age group and varied for the middle and old age groups. Generally, XGBoost and random forest models outperformed other models. However, the XGBoost models were slightly better than the random forest models, except for the young age group.
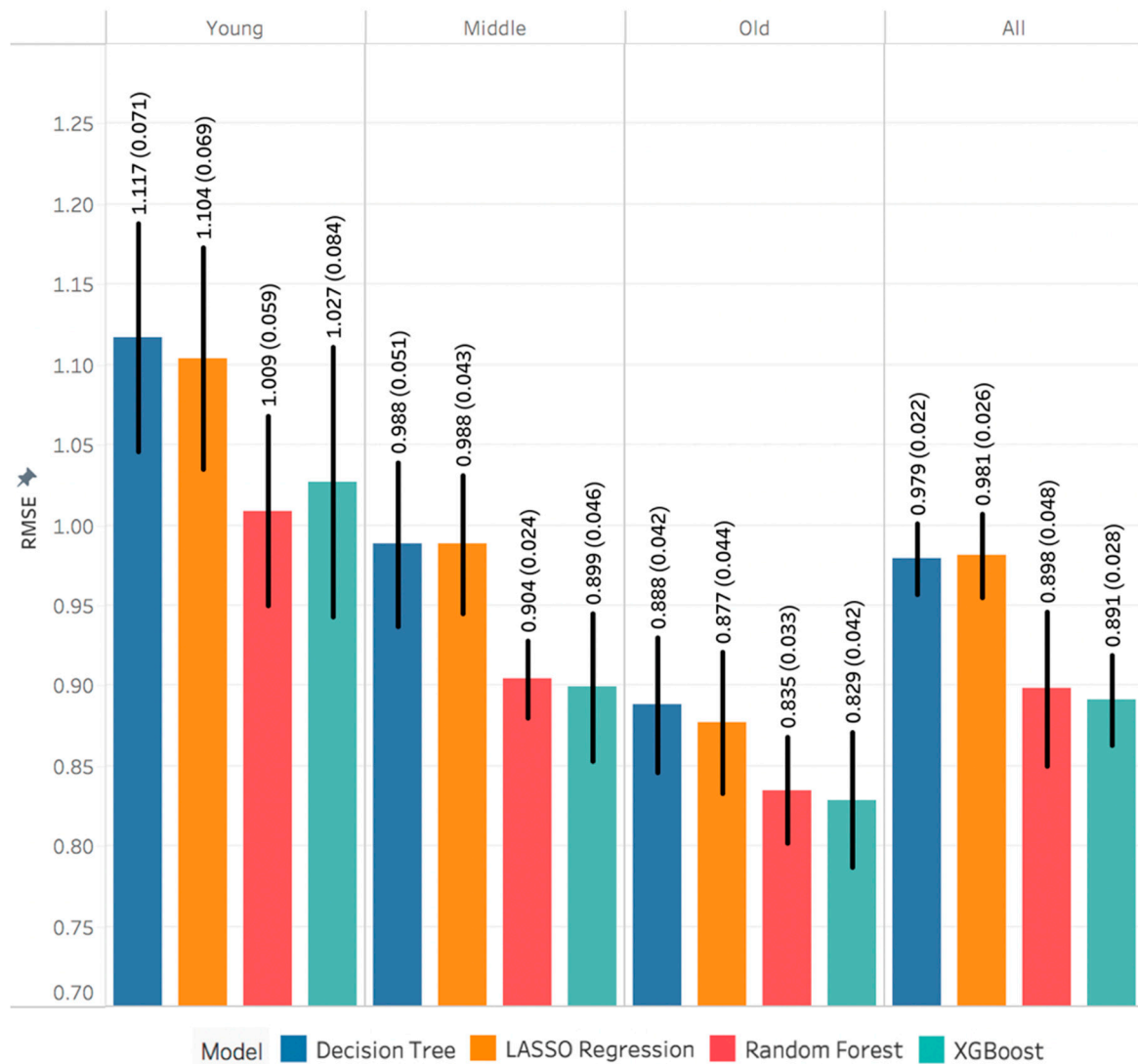

Figure 4. Performance metrics of estimating energy expenditure. Each value is the mean and standard deviation of the 5-fold nested cross validation.

Table 3 shows the performance of recognizing individual PA using XGBoost. 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.

Table 3. Performance metrics of recognizing individual physical activities using XGBoost. 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	0.544 (0.055)	0.491 (0.070)	0.391 (0.059)	0.497 (0.026)
rapid walk	0.645 (0.055)	0.545 (0.061)	0.470 (0.048)	0.567 (0.037)
light gardening	0.585 (0.056)	0.529 (0.051)	0.495 (0.025)	0.571 (0.047)
yard work	0.416 (0.035)	0.478 (0.046)	0.404 (0.070)	0.489 (0.040)
prepare serve meal	0.520 (0.022)	0.482 (0.037)	0.480 (0.046)	0.520 (0.027)
digging	0.711 (0.040)	0.686 (0.050)	0.637 (0.053)	0.719 (0.038)
straightening up dusting	0.460 (0.051)	0.427 (0.041)	0.415 (0.027)	0.483 (0.014)
washing dishes	0.782 (0.012)	0.706 (0.024)	0.596 (0.035)	0.716 (0.023)
unloading storing dishes	0.666 (0.031)	0.669 (0.044)	0.597 (0.036)	0.675 (0.021)
walking at rpe 1	0.366 (0.064)	0.491 (0.027)	0.318 (0.056)	0.437 (0.027)
personal care	0.660 (0.043)	0.709 (0.028)	0.552 (0.027)	0.672 (0.011)
dressing	0.494 (0.035)	0.450 (0.038)	0.335 (0.023)	0.456 (0.021)
walking at rpe 5	0.482 (0.050)	0.440 (0.104)	0.356 (0.094)	0.443 (0.029)
sweeping	0.602 (0.068)	0.634 (0.073)	0.518 (0.057)	0.625 (0.018)
vacuuming	0.637 (0.029)	0.611 (0.044)	0.533 (0.035)	0.625 (0.024)
stair descent	0.705 (0.120)	0.693 (0.055)	0.635 (0.064)	0.706 (0.040)
stair ascent	0.543 (0.104)	0.561 (0.085)	0.518 (0.023)	0.600 (0.047)
trash removal	0.425 (0.047)	0.473 (0.050)	0.355 (0.017)	0.465 (0.034)
replacing sheets on a bed	0.626 (0.064)	0.677 (0.029)	0.559 (0.024)	0.665 (0.031)
stretching yoga	0.628 (0.026)	0.642 (0.033)	0.557 (0.043)	0.630 (0.035)
mopping	0.673 (0.039)	0.660 (0.033)	0.623 (0.041)	0.702 (0.041)
light home maintenance	0.507 (0.027)	0.536 (0.035)	0.459 (0.028)	0.530 (0.025)
computer work	0.780 (0.043)	0.800 (0.039)	0.759 (0.049)	0.795 (0.017)
heavy lifting	0.650 (0.031)	0.672 (0.024)	0.495 (0.041)	0.647 (0.035)
shopping	0.506 (0.052)	0.537 (0.039)	0.524 (0.033)	0.563 (0.040)
ironing	0.636 (0.033)	0.687 (0.014)	0.620 (0.056)	0.700 (0.023)
laundry washing	0.426 (0.036)	0.509 (0.039)	0.411 (0.040)	0.479 (0.021)
strength exercise leg curl	0.576 (0.044)	0.644 (0.062)	0.656 (0.108)	0.695 (0.028)
strength exercise chest press	0.681 (0.082)	0.668 (0.063)	0.602 (0.079)	0.697 (0.017)
strength exercise leg extension	0.367 (0.128)	0.462 (0.092)	0.329 (0.079)	0.419 (0.019)
tv watching	0.614 (0.050)	0.616 (0.019)	0.546 (0.069)	0.624 (0.030)
standing still	0.631 (0.081)	0.644 (0.060)	0.527 (0.094)	0.612 (0.036)
washing windows	0.764 (0.058)	0.720 (0.045)	0.739 (0.056)	0.755 (0.024)
Macro average (F1 score)	0.584 (0.023)	0.594 (0.021)	0.516 (0.011)	0.600 (0.014)

Figures S7–S9 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 Figure 2. Figures S10–S12 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 Figure 4.

Figures S13–S15 show the top 15 features that contributed the most in recognizing PA type across age groups extracted from the XGBoost models. It can be noticed that the ranking of features is similar across age groups within each PA type. Figures S16–S18 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.

4. Discussion

The goal of the study was to build accurate machine learning models to recognize the hallmark measures of physical activities and estimate energy expenditure across different age groups. We analyzed a large dataset of raw accelerometer data collected from the wrist position. We utilized four machine learning algorithms to build our models including: decision tree, random forest, extreme gradient boosting (XGBoost) and least absolute shrinkage

and selection operator (LASSO). 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 Figure 2. Although the results were similar across age groups, there was a slightly higher performance in the young, followed by the middle, then the old age group for a majority of the 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 S7–S9. 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, which occurred more commonly among the older. An additional cause could be that the middle and old age groups include more participant 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 S7–S9.

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 Figure 3. The highest performance was for the young and middle age groups alternatively and then the 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 Figure 3 and Figures S10–S12). If this error was scaled to free-living conditions over a typical day (16 h), older adults would be expected to have 2% (~19 min) 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 0.8 for recognizing computer work activity in the middle age group and the lowest was 0.318 for recognizing walking at RPE 1 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 Figures 2 and 3. In general, there were no consistent differences among age groups.

The scaled impurity-based feature importance ranking generated from the XGBoost 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 generated from the XGBoost models. 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 were 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 regarding 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, which 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. [49] 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. Despite the differences mentioned above, we compared our work with others who processed data collected from a triaxial accelerometer placed on the wrist position to recognize physical activity type, recognize individual activities or estimate energy expenditure, as shown in the Supplemental Table S2.

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) [50], PAMAP2 (chest, arm and ankle positions, 9 participants) [51], UCI daily and sports dataset (hip position, 30 participants) [52], Skoda Mini Checkpoint (multiple body positions, 1 participant) [53], WISDM (hip position, 29 participants) [54], and Daphnet Freezing of Gait Dataset (legs and hip positions, 10 participants) [52]). They are also limited by a small number of participants, the 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 models 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 [55]. Collecting data in free-living settings is more reflective of the 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 performance of machine learning models in estimating activity types, activity intensity and energy expenditure would vary across age groups. Overall results suggest data features derived from wrist worn accelerometers and analyzed using machine learning models lead to high-to-moderate accuracy across all age groups. In conclusion, a generalized approach to processing wrist accelerometry data, without consideration of a person's age, is sufficient for estimating physical activity.

References

1. Guthold, R.; Stevens, G.A.; Riley, L.M.; Bull, F.C. 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. Health* **2018**, *6*, e1077–e1086. [CrossRef]
2. Branch, L.G.; Jette, A.M. A prospective study of long-term care institutionalization among the aged. *Am. J. Public Health* **1982**, *72*, 1373–1379. [CrossRef] [PubMed]
3. Corti, M.-C.; Guralnik, J.M.; Salive, M.E.; Ferrucci, L.; Pahor, M.; Wallace, R.B.; Hennekens, C.H. Serum Iron Level, Coronary Artery Disease, and All-Cause Mortality in Older Men and Women. *Am. J. Cardiol.* **1997**, *79*, 120–127. [CrossRef]
4. Khokhar, S.R.; Stern, Y.; Bell, K.; Anderson, K.; Noe, E.; Mayeux, R.; Albert, S.M. Persistent Mobility Deficit in the Absence of Deficits in Activities of Daily Living: A Risk Factor for Mortality. *J. Am. Geriatr. Soc.* **2001**, *49*, 1539–1543. [CrossRef]
5. Newman, A.B.; Simonsick, E.M.; Naydeck, B.L.; Boudreau, R.M.; Kritchevsky, S.B.; Nevitt, M.C.; Pahor, M.; Satterfield, S.; Brach, J.S.; Studenski, S.A.; et al. Association of Long-Distance Corridor Walk Performance With Mortality, Cardiovascular Disease, Mobility Limitation, and Disability. *JAMA* **2006**, *295*, 2018–2026. [CrossRef] [PubMed]
6. NCDs|WHO Launches ACTIVE: A Toolkit for Countries to Increase Physical Activity and Reduce Noncommunicable Diseases, WHO. 2018. Available online: http://www.who.int/ncds/prevention/physical-activity/active-toolkit/en/ (accessed on 9 December 2020).
7. WHO. Guidelines on Physical Activity and Sedentary Behaviour. Available online: https://www.who.int/publications/i/item/9789240015128 (accessed on 15 January 2021).
8. Staudenmayer, J.; He, S.; Hickey, A.; Sasaki, J.E.; Freedson, P.S. Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *J. Appl. Physiol.* **2015**, *119*, 396–403. [CrossRef]
9. Eellis, K.; Egodbole, S.; Emarshall, S.; Elanckriet, G.; Estaudenmayer, J.; Ekerr, J. Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms. *Front. Public Health* **2014**, *2*, 36. [CrossRef] [PubMed]
10. Kheirkhahan, M.; Mehta, S.; Nath, M.; Wanigatunga, A.A.; Corbett, D.B.; Manini, T.M.; Ranka, S. A Bag-of-Words Approach for Assessing Activities of Daily Living Using Wrist Accelerometer Data. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas, MO, USA, 13–16 January 2017; pp. 678–685.
11. Nakandala, S.; Jankowska, M.M.; Tuz-Zahra, F.; Bellettiere, J.; Carlson, J.A.; LaCroix, A.Z.; Hartman, S.J.; Rosenberg, D.E.; Zou, J.; Kumar, A.; et al. Application of Convolutional Neural Network Algorithms for Advancing Sedentary and Activity Bout Classification. *J. Meas. Phys. Behav.* **2020**, *1*, 1–9. [CrossRef]
12. Mehta, P.; Bukov, M.; Wang, C.-H.; Day, A.G.; Richardson, C.; Fisher, C.K.; Schwab, D.J. A high-bias, low-variance introduction to Machine Learning for physicists. *Phys. Rep.* **2019**, *810*, 1–124. [CrossRef] [PubMed]
13. Khan, M.A.; Karim, R.; Kim, Y. A Two-Stage Big Data Analytics Framework with Real World Applications Using Spark Machine Learning and Long Short-Term Memory Network. *Symmetry* **2018**, *10*, 485. [CrossRef]
14. Ntampaka, M.; Trac, H.; Sutherland, D.J.; Fromenteau, S.; Póczos, B.; Schneider, J. Dynamical Mass Measurements of Contaminated Galaxy Clusters Using Machine Learning. *Astrophys. J.* **2016**, *831*, 135. [CrossRef]
15. Dijkhuis, T.B.; Blaauw, F.J.; Van Ittersum, M.W.; Velthuijsen, H.; Aiello, M. Personalized Physical Activity Coaching: A Machine Learning Approach. *Sensors* **2018**, *18*, 623. [CrossRef] [PubMed]
16. Kańtoch, E. Recognition of Sedentary Behavior by Machine Learning Analysis of Wearable Sensors during Activities of Daily Living for Telemedical Assessment of Cardiovascular Risk. *Sensors* **2018**, *18*, 3219. [CrossRef] [PubMed]
17. Chowdhury, A.K.; Tjondronegoro, D.; Chandran, V.; Zhang, J.; Trost, S.G. Prediction of Relative Physical Activity Intensity Using Multimodal Sensing of Physiological Data. *Sensors* **2019**, *19*, 4509. [CrossRef]
18. Javed, A.R.; Sarwar, M.U.; Khan, S.; Iwendi, C.; Mittal, M.; Kumar, N. Analyzing the Effectiveness and Contribution of Each Axis of Tri-Axial Accelerometer Sensor for Accurate Activity Recognition. *Sensors* **2020**, *20*, 2216. [CrossRef]
19. Ahmadi, M.N.; Pavey, T.G.; Trost, S.G. Machine Learning Models for Classifying Physical Activity in Free-Living Preschool Children. *Sensors* **2020**, *20*, 4364. [CrossRef] [PubMed]
20. Attal, F.; Mohammed, S.; Dedabrishvili, M.; Chamroukhi, F.; Oukhellou, L.; Amirat, Y. Physical Human Activity Recognition Using Wearable Sensors. *Sensors* **2015**, *15*, 31314–31338. [CrossRef]
21. Troiano, R.P.; McClain, J.J.; Brychta, R.J.; Chen, K.Y. Evolution of accelerometer methods for physical activity research. *Br. J. Sports Med.* **2014**, *48*, 1019–1023. [CrossRef] [PubMed]
22. Migueles, J.H.; Cadenas-Sanchez, C.; Ekelund, U.; Nyström, C.D.; Mora-Gonzalez, J.; Löf, M.; Labayen, I.; Ruiz, J.R.; Ortega, F.B. Accelerometer Data Collection and Processing Criteria to Assess Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations. *Sports Med.* **2017**, *47*, 1821–1845. [CrossRef]
23. Kerr, J.; Marinac, C.R.; Ellis, K.; Godbole, S.; Hipp, A.; Glanz, K.; Mitchell, J.; Laden, F.; James, P.; Berrigan, D. Comparison of Accelerometry Methods for Estimating Physical Activity. *Med. Sci. Sports Exerc.* **2017**, *49*, 617–624. [CrossRef]
24. Worldwide Wearables Market Forecast to Maintain Double-Digit Growth in 2020 and Through 2024, According to IDC. Available online: https://www.idc.com/getdoc.jsp?containerId=prUS46885820 (accessed on 9 December 2020).
25. Full, K.M.; Kerr, J.; Grandner, M.A.; Malhotra, A.; Moran, K.; Godoble, S.; Natarajan, L.; Soler, X. Validation of a physical activity accelerometer device worn on the hip and wrist against polysomnography. *Sleep Health* **2018**, *4*, 209–216. [CrossRef] [PubMed]

26. Kinnunen, H.; Häkkinen, K.; Schumann, M.; Karavirta, L.; Westerterp, K.R.; Kyröläinen, H. Training-induced changes in daily energy expenditure: Methodological evaluation using wrist-worn accelerometer, heart rate monitor, and doubly labeled water technique. *PLoS ONE* **2019**, *14*, e0219563. [CrossRef] [PubMed]

27. O'Driscoll, R.; Turicchi, J.; Beaulieu, K.; Scott, S.; Matu, J.; Deighton, K.; Finlayson, G.; Stubbs, J. How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. *Br. J. Sports Med.* **2018**, *54*, 332–340. [CrossRef]

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

29. Knaggs, J.D.; Larkin, K.A.; Manini, T.M. Metabolic Cost of Daily Activities and Effect of Mobility Impairment in Older Adults. *J. Am. Geriatr. Soc.* **2011**, *59*, 2118–2123. [CrossRef]

30. American Time Use Survey Home Page, (n.d.). Available online: https://www.bls.gov/tus/ (accessed on 9 December 2020).

31. Jetté, M.; Sidney, K.; Blümchen, G. Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. *Clin. Cardiol.* **1990**, *13*, 555–565. [CrossRef]

32. Krause, A.; Siewiorek, D.; Smailagic, A.; Farringdon, J. Unsupervised, Dynamic Identification of Physiological and Activity Context in Wearable Computing. In Proceedings of the Seventh IEEE International Symposium on Wearable Computers, White Plains, NY, USA, 21–23 October 2003; pp. 88–97.

33. Mannini, A.; Intille, S.S.; Rosenberger, M.; Sabatini, A.M.; Haskell, W. Activity Recognition Using a Single Accelerometer Placed at the Wrist or Ankle. *Med. Sci. Sports Exerc.* **2013**, *45*, 2193–2203. [CrossRef] [PubMed]

34. Stikic, M.; Huynh, T.; Van Laerhoven, K.; Schiele, B. ADL recognition based on the combination of RFID and accelerometer sensing. In Proceedings of the 2008 Second International Conference on Pervasive Computing Technologies for Healthcare, Tampere, Finland, 30 January–1 February 2008; pp. 258–263.

35. Huynh, T.; Schiele, B. *Analyzing Features for Activity Recognition*; Association for Computing Machinery (ACM): New York, NY, USA, 2005; pp. 159–163.

36. Karantonis, D.; Narayanan, M.; Mathie, M.; Lovell, N.; Celler, B. Implementation of a Real-Time Human Movement Classifier Using a Triaxial Accelerometer for Ambulatory Monitoring. *IEEE Trans. Inf. Technol. Biomed.* **2006**, *10*, 156–167. [CrossRef]

37. Pirttikangas, S.; Fujinami, K.; Nakajima, T. Feature Selection and Activity Recognition from Wearable Sensors. In *Lecture Notes in Computer Science*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2006; pp. 516–527.

38. Ermes, M.; Pärkkä, J.; Mäntyjärvi, J.; Korhonen, I. Detection of Daily Activities and Sports with Wearable Sensors in Controlled and Uncontrolled Conditions. *IEEE Trans. Inf. Technol. Biomed.* **2008**, *12*, 20–26. [CrossRef]

39. Davoudi, A.; Wanigatunga, A.A.; Kheirkhahan, M.; Corbett, D.B.; Mendoza, T.; Battula, M.; Ranka, S.; Fillingim, R.B.; Manini, T.M.; Rashidi, P. Accuracy of Samsung Gear S Smartwatch for Activity Recognition: Validation Study. *JMIR mHealth uHealth* **2019**, *7*, e11270. [CrossRef] [PubMed]

40. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.

41. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [CrossRef]

42. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]

43. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*, 1st ed.; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-6848-6.

44. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCIA* **1995**, *14*, 1137–1143.

45. Molinaro, A.M.; Simon, R.; Pfeiffer, R.M. Prediction error estimation: A comparison of resampling methods. *Bioinformatics* **2005**, *21*, 3301–3307. [CrossRef]

46. Cawley, G.C.; Talbot, N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.

47. Friedman, J.H.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]

48. Rokach, L.; Maimon, O. *Data Mining with Decision Trees*; World Scientific Publishing: Singapore, 2013.

49. Ellis, K.; Kerr, J.; Godbole, S.; Lanckriet, G.; Wing, D.; Marshall, S. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiol. Meas.* **2014**, *35*, 2191–2203. [CrossRef] [PubMed]

50. Chavarriaga, R.; Sagha, H.; Calatroni, A.; Digumarti, S.T.; Tröster, G.; Millán, J.D.R.; Roggen, D. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognit. Lett.* **2013**, *34*, 2033–2042. [CrossRef]

51. Reiss, A.; Stricker, D. Introducing a New Benchmarked Dataset for Activity Monitoring. In Proceedings of the 16th International Symposium on Wearable Computers, Seattle, WA, USA, 7–10 October 2012; pp. 108–109. [CrossRef]

52. Asuncion, A.; Newman, D.J. *UCI Machine Learning Repository*; University of California, School of Information and Computer Science: Irvine, CA, USA, 2007; Available online: http://archive.ics.uci.edu/ml/index.php (accessed on 2 October 2018).

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

54. WISDM Lab: Dataset, (n.d.). Available online: https://www.cis.fordham.edu/wisdm/dataset.php (accessed on 13 January 2021).

55. Keadle, S.K.; Lyden, K.A.; Strath, S.J.; Staudenmayer, J.W.; Freedson, P.S. A Framework to Evaluate Devices That Assess Physical Behavior. *Exerc. Sport Sci. Rev.* **2019**, *47*, 206–214. [CrossRef]
56. Zhang, S.; Rowlands, A.V.; Murray, P.; Hurst, T.L. Physical Activity Classification Using the GENEA Wrist-Worn Accelerometer. *Med. Sci. Sports Exerc.* **2012**, *44*, 742–748. [CrossRef]
57. Trost, S.G.; Zheng, Y.; Wong, W.-K. Machine learning for activity recognition: Hip versus wrist data. *Physiol. Meas.* **2014**, *35*, 2183–2189. [CrossRef]