




Article

# Machine Learning Models for Classifying Physical Activity in Free-Living Preschool Children

Matthew N. Ahmadi <sup>1,2</sup>, Toby G. Pavey <sup>2</sup> and Stewart G. Trost <sup>1,2,\*</sup>

<sup>1</sup> Institute of Health and Biomedical Innovation at Queensland Centre for Children’s Health Research, Queensland University of Technology, South Brisbane 4101, Australia; matthewnguyen.ahmadi@hdr.qut.edu.au

<sup>2</sup> Faculty of Health, School of Exercise and Nutrition Sciences, Queensland University of Technology, Kelvin Grove 4059, Australia; toby.pavey@qut.edu.au

\* Correspondence: s.trost@qut.edu.au; Tel.: +61-7-3069-7301

Received: 26 June 2020; Accepted: 4 August 2020; Published: 5 August 2020



**Abstract:** Machine learning (ML) activity classification models trained on laboratory-based activity trials exhibit low accuracy under free-living conditions. Training new models on free-living accelerometer data, reducing the number of prediction windows comprised of multiple activity types by using shorter windows, including temporal features such as standard deviation in lag and lead windows, and using multiple sensors may improve the classification accuracy under free-living conditions. The objective of this study was to evaluate the accuracy of Random Forest (RF) activity classification models for preschool-aged children trained on free-living accelerometer data. Thirty-one children (mean age = 4.0 ± 0.9 years) completed a 20 min free-play session while wearing an accelerometer on their right hip and non-dominant wrist. Video-based direct observation was used to categorize the children’s movement behaviors into five activity classes. The models were trained using prediction windows of 1, 5, 10, and 15 s, with and without temporal features. The models were evaluated using leave-one-subject-out-cross-validation. The F-scores improved as the window size increased from 1 to 15 s (62.6%–86.4%), with only minimal improvements beyond the 10 s windows. The inclusion of temporal features increased the accuracy, mainly for the wrist classification models, by an average of 6.2 percentage points. The hip and combined hip and wrist classification models provided comparable accuracy; however, both the models outperformed the models trained on wrist data by 7.9 to 8.2 percentage points. RF activity classification models trained with free-living accelerometer data provide accurate recognition of young children’s movement behaviors under real-world conditions.

**Keywords:** physical activity; accelerometer; measurement; supervised learning; classification; assessment; early childhood

## 1. Introduction

Childhood obesity continues to be a serious global public health problem. In 2016, more than 41 million children between 0 and 5 years were overweight or obese [1]. The high prevalence of overweight and obesity among young children is cause for concern. Preschool-aged children who are overweight or obese are at an increased risk of type 2 diabetes, cardiovascular disease, and mental health problems as they progress to pre-adolescence and adulthood [2,3]. Childhood obesity is the top contributor to health care cost across all decades of life, with a 3 to 5 times higher health-care cost burden in adults with a history of childhood obesity [4].

Physical inactivity is a modifiable risk factor that contributes to the increase in the prevalence of overweight and obesity observed in young children [5,6]. Consequently, public health authorities

have identified the preschool years as a critical period to intervene and promote regular physical activity [7–9]. The accurate measurement of physical activity is essential in order to identify and understand the individual, environmental and sociocultural determinants of physical activity, and to evaluate the effectiveness of physical activity intervention programs. Due to their unobtrusive size, robustness, and low cost, accelerometer-based motion sensors have become the method of choice for measuring physical activity in studies involving preschool-aged children [10].

Traditionally, cut-point methods have been used to classify physical activity intensity and estimate the time spent sedentary and in light, moderate, and vigorous physical activity. With this approach, the relationship between the measured energy expenditure and accelerometer counts is established using linear regression and thresholds or “cut-points” denoting the dividing line between sedentary-and-light (1.5 Metabolic Equivalents (METs)), light-and-moderate (3–4 METs), and moderate-and-vigorous physical activity (6 METs). Another common cut-point approach is the use of receiver operating characteristic (ROC) curves to determine the count threshold that provides the optimal combination of sensitivity and specificity for distinguishing between adjacent levels of physical activity intensity [11,12]. However, cut-point methods have been shown to have high misclassification rates among children. Validation studies involving independent samples of children indicate that cut-point approaches misclassify the true intensity of physical activity 35% to 45% of the time [13–16].

An alternative to cut-point methods is pattern recognition approaches, such as machine learning. Pattern recognition is a branch of artificial intelligence concerned with classifying or describing observations, with the goal of predicting outcomes based on previous knowledge or recognizable features in the raw data [17–19]. Accelerometer data processing techniques based on pattern recognition have been shown to provide accurate predictions of physical activity type and more accurate assessments of physical activity intensity [20–22]. Nonetheless, the uptake of machine learning methods by physical activity researchers has been slow, in part due to the difficulties of implementation, and the consistent finding that models trained on accelerometer data from laboratory-based activity trials do not generalize well to free-living environments [23–25].

We recently evaluated the accuracy of laboratory-trained machine learning activity classification models for preschool-aged children under true free-living conditions [26]. The models classified children’s physical activity behaviors activities into one of five activity classes: sedentary, light activities and games, moderate to vigorous activities and games, walk, and run. Under laboratory conditions, the overall classification accuracy ranged between 80% and 82% for models trained on hip data and 78% to 81% for models trained on wrist accelerometer data. However, when evaluated under true free-living conditions, the overall accuracy decreased by 10 to 20 percentage points to 66% to 70% for the hip models and 59% to 60% for the wrist models. Notably, there were substantial decreases in the walking recognition accuracy, which ranged from just 8% to 11% for the hip and 12% to 15% for the wrist. The reductions in classification accuracy under free-living conditions were attributed to several important methodological limitations. First, the laboratory-based models were trained using accelerometer data from a limited number of activities performed in a standardized manner. Under free-living conditions, a wide range of activities are performed, and physical activity behavior is far more variable. Second, the 15 s prediction window employed by the models may have been too long to capture the pulsatile and sporadic activity behaviors of preschool-aged children, resulting in prediction windows comprised of multiple activity types that are more difficult to classify. Third, the predictions for each activity window were made without considering temporal features, such as the variability in accelerometer signal in the preceding and succeeding activity windows. Incorporating information from the lag and lead windows may reduce noise and improve the classification of physical activity sequences. Fourth and finally, the classification models did not utilize features from multiple sensor placements through feature fusion. Models trained on data from multiple accelerometers placements have the potential to mitigate the weaknesses a single monitor placement may have for the detection of certain activities, such as activities with extensive upper body movement. While the aforementioned limitations of laboratory-based models have been reported previously [26], to our knowledge no

previous study has developed activity classification models for preschool children trained exclusively on free-living data.

With this in mind, the current study evaluated the accuracy of machine learning activity classification models for preschool-aged children trained on true free-living accelerometer data. To address the limitations identified in our recent evaluation of laboratory-trained models, the classifiers were trained over a range of window sizes ranging from 1 to 15 s, the classification accuracy was evaluated with and without the inclusion of temporal features, and the models based on multiple accelerometer placements and feature fusion were evaluated. We hypothesized that the use of smaller prediction windows, the inclusion of temporal features, and feature fusion from multiple accelerometer placements would increase the classification performance under free-living conditions.

## 2. Materials and Methods

### 2.1. Participants

A total of 31 children between the ages of 3 and 5 years participated in the study. The sample was comprised of 9 girls and 22 boys, and there were approximately 10 children in each age category [26]. The children were recruited through a university email list-serv, local media, and word of mouth. Written parental consent was obtained prior to participation. The study was approved by the Queensland University of Technology's Human Research Ethics Committee (approval number: 1700000423; date of approval—July 14, 2017).

### 2.2. Free-Living Play Session

Each child completed a 20 min active free play session at a location chosen by their parent or guardian. The locations that were chosen included the family home, community parks, and local green spaces [26]. The research team provided age-appropriate toys and play equipment, and the children were free to engage in any activity they desired. This allowed for natural activity behavior, transitions, and engagement with peers and the environment. The children were video recorded during the free play sessions with a hand-held Go-Pro Hero 5 (GoPro, Inc., San Mateo, CA, USA) camera for subsequent direct observation coding. Prior to each play session, an external timepiece was synchronized with the laptop computer used to initialize the accelerometers and displayed in front of the camera to ensure synchronization between the Go-Pro video files and accelerometer timestamps [26].

### 2.3. Instrumentation

During each free play session, the children wore an ActiGraph GT3X+ accelerometer (ActiGraph Corporation; Pensacola, FL, USA). The ActiGraph GT3X+ is a small and lightweight monitor that measures acceleration along three orthogonal axes with a dynamic range between  $\pm 6$  g and a sampling frequency between 30 and 100 Hz. For the current study, the sampling frequency was set to 100 Hz. The ActiGraph monitors were worn on the children's right hip and non-dominant wrist. For the hip location, the accelerometer was positioned on the right mid-axilla line at the level of the iliac crest. For the wrist location, the accelerometer was positioned on the posterior side of the arm, between the radial and ulnar styloid processes.

### 2.4. Direct Observation Coding Procedure

Go-Pro video files were imported into the Noldus Observer XT 14 software (Noldus Information Technology, Wageningen, The Netherlands) for continuous direct observation coding. A customized direct observation scheme was implemented in which the participant's movement behavior was coded as one of the five activity classes predicted by the activity classification models [26]: sedentary (SED), light activities and games (LIGHT\_AG), moderate-vigorous activities and games (MV\_AG), walking (WALK), and running (RUN). A description of the activity classes is provided in Table 1. If a participant was not in view of the camera, movement behavior was coded as "out of view". The computerized

direct observation system generated a vector of date-time stamps corresponding to the start and finish of each movement event, which were used to calculate the event duration and assign the activity codes to the corresponding time segments of the accelerometer data. The inter-observer reliability was assessed by having two researchers independently code five randomly selected videos. Cohen’s unweighted kappa statistic for activity class was 0.86 (95% CI: 0.84–0.88), which, according to the ratings suggested by Landis and Koch [27], is almost perfect agreement.

**Table 1.** Description of the five activity classes.

Activity Class	Movement Descriptors	Activity Types
SED	Sitting/lying down Stationary/motionless	Sit still Sit w/upper body movement
LIGHT_AG	Standing Stationary/movement of limbs or trunk (very easy) Translocation (slow/easy)	Stand still Stand w/upper body movement Crawl Up/downstairs Floor games Stand and kick Slide Climb (low intensity)
MV_AG	Translocation (medium speed/moderate) Translocation (fast or very fast/hard)	Run and kick Side gallop Jump/hop/leap Ride a bike Ride a scooter Stationary ride/spin/swing Climb (high intensity)
WALK	Translocation (steady/medium speed/moderate)	Walk slow/stroll Walk brisk Walk and hold object
RUN	Translocation (steady/fast or very fast/hard)	Sprint Run and hold object

SED = sedentary; LIGHT AG = light activity and games; MV AG = moderate to vigorous activities and games; WALK = walking; RUN = running

## 2.5. Development and Evaluation of Activity Classification Models

### 2.5.1. Data Processing and Feature Extraction

The annotated accelerometer data were segmented into non-overlapping sliding windows of 1, 5, 10, and 15 s. If the window contained multiple activity class codes (i.e., a combination of walking and running), the assigned code represented the activity class completed for the majority of the window (>50% window duration). To determine the impact of these “mixed windows” on the classifier performance, an indicator variable was created to flag windows with multiple activity codes. This enabled us to determine if the window size influenced the classifier performance by reducing the number of mixed prediction windows. Within each window, the tri-axial accelerometer signal was transformed into a single-dimension vector magnitude (VM), and two sets of features were extracted: the VM was selected to match the laboratory-based classification models and reduce the dimensionality.

- (1) Base features: time and frequency domain features were used in the previously published activity classification models [21,22,28]: mean, SD, minimum, maximum, interquartile range, percentiles (10th, 25th, 50th, 75th, 95th), coefficient of variation, signal sum, signal power, peak-to-peak amplitude, median crossings, cross axis correlations, dominant frequency between 0.25 and 5.0 Hz, and magnitude of dominant frequency between 0.25 and 5.0 Hz.

- (2) Base plus temporal features: a second feature set consisted of the base features and temporal features calculated from the preceding (lead) and succeeding (lag) activity windows. These included the standard deviation (SD) for the 1 and 2 lag and lead windows and the SD over the lag and lead windows and the current window ( $\sigma = \frac{1}{N} \sum_{i=0}^N (x_i - \mu)^2$ , where  $n = 5$ ). This resulted in a total of five temporal features, in addition to the base features.

### 2.5.2. Model Training and Testing

Random Forest (RF) classification models were developed for each placement. An RF model was chosen because: (1) it is a widely implemented ensemble-based supervised learning algorithm that has been shown to provide robust results in prior activity recognition studies [29]; (2) it requires minimal data pre-processing, as the features do not need to be normalized; and (3) feature selection procedures are not required because the algorithm effectively does this on its own [30]. The models were implemented using the “randomForest” and “caret” packages within R software. The number of features randomly sampled at each node (tuning parameter mtry) was optimized during training, whilst the number of trees was kept constant at 500. The fully annotated training datasets, r scripts for the final RF models, and data required to run the classification models are available at the link provided in the Supplementary Materials section.

The out of sample prediction error was evaluated using leave-one-subject-out-cross-validation (LOSO-CV). In LOSO-CV, the model is trained on data from all the participants except one, which is “held out” and used as the test data set. The process is repeated until all the participants have served as the test data, and the performance metrics are aggregated [31]. F-scores were used to assess the accuracy of each model. F-scores were calculated because they are based on the harmonic mean of precision and recall and are less biased by class size imbalances [32]. The F-scores were calculated for each activity class and averaged to provide an overall F-score for each classifier. Additionally, to identify patterns of misclassification, confusion matrices were generated for each model.

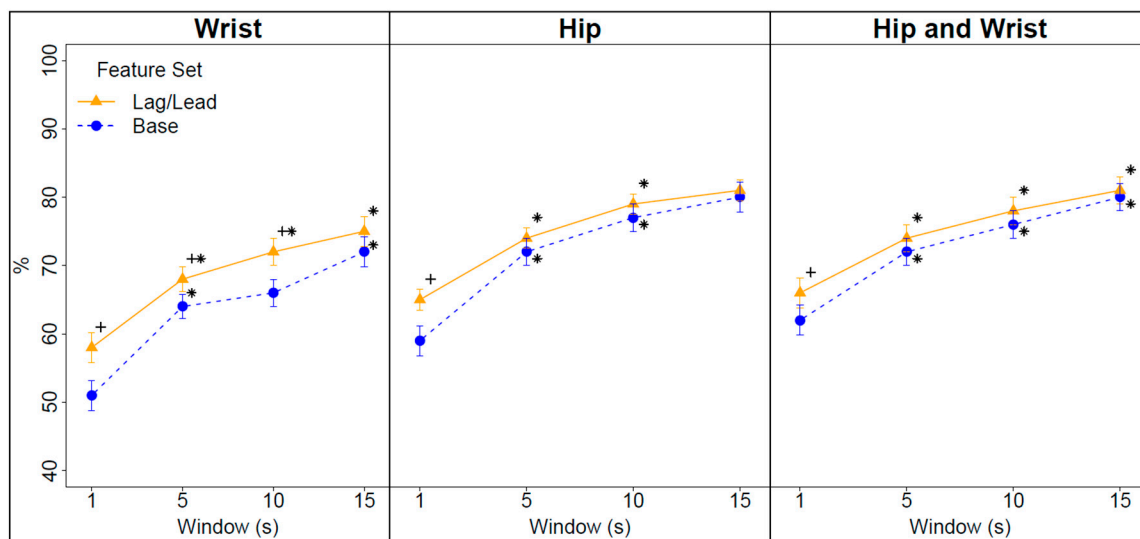
### 2.6. Statistical Analysis

A  $3 \times 2 \times 4$  repeated measures ANOVA was run to examine the effects of placement (wrist, hip, combined wrist and hip), feature set (base vs. temporal features), and window size (1, 5, 10, and 15 s) on the F-scores. The significant main effects and interactions were evaluated using tests of simple effects and pre-planned single degree freedom contrasts. Statistical significance was set at an alpha level of 0.05.

## 3. Results

The results of the repeated measures ANOVA identified a significant two-way feature set by window size interaction ( $F_{3,90} = 3.40, p = 0.02$ ), indicating that the effects of the feature set on the F-scores differed by window size. The two-way placement by feature set interaction ( $F_{2,60} = 1.88, p = 0.17$ ), placement by window size interaction ( $F_{6,180} = 0.79, p = 0.58$ ), and three-way placement by feature set by window size interaction ( $F_{6,180} = 0.20, p = 0.97$ ) were not statistically significant.

Figure 1 displays interaction plots summarizing the effects of window size and feature set on the F-scores at each accelerometer placement. The numbers of prediction intervals with multiple activity types (mixed windows) for models trained on 1, 5, 10, and 15 s windows were 3527 (9.2%), 2255 (29.5%), 1597 (42.1%), and 1258 (50.2%), respectively. For the wrist and combined hip and wrist models, the F-scores increased significantly as the window size increased from 1 to 15 s. However, for the models trained on hip data, the F-scores failed to improve significantly as the window size increased from 10 to 15 s.



**Figure 1.** Interaction plots summarizing the effect of window size and feature set on the adjusted F-scores for models trained on wrist, hip, and combined hip and wrist accelerometer data. + Denotes significantly different from the base model at a given window size  $p < 0.05$ ; \* Denotes significantly different from the previous window size for a given feature set  $p < 0.05$ .

For the models trained on wrist data on 1, 5, and 10 s windows and the models trained on the hip and combined hip and wrist data on 1 s windows, the addition of lag/lead features resulted in small but significant improvements in the F-scores. For the wrist model trained on 15 s windows and the hip and combined hip and wrist models trained on 5, 10, and 15 s windows, the addition of lag/lead features did not significantly improve the performance.

Across all window sizes and for both feature sets, the classification models trained on the wrist accelerometer data exhibited significantly lower F-scores than the models trained on the hip or combined hip and wrist accelerometer data. On average, the performance differential was 7.9 to 8.2 percentage points. There were no significant differences in the F-scores between the models trained on the hip accelerometer data and the combined hip and wrist accelerometer data.

The F-scores for the five activity classes and the weighted average F-score for each model are reported in Table 2. The class-level F-scores tended to increase as the window size increased from 1 to 15 s. The inclusion of lag/lead features resulted in marginal increases in the F-scores; however, this was not true for all models, particularly for the WALK and RUN activity classes. Only the lag/lead wrist model trained on 10 s windows and the hip and combined hip and wrist models trained on the 10 and 15 s windows exhibited an average F-score of  $\geq 80\%$  and provided F-scores of  $\geq 70\%$  for all five activity classes. The wrist lag/lead model trained on 15 s windows achieved an average F-score of  $\geq 80\%$ ; however, the F-score for WALK was just under 70%.

**Table 2.** F-scores for the five activity classes and the weighted average F-score for each model.

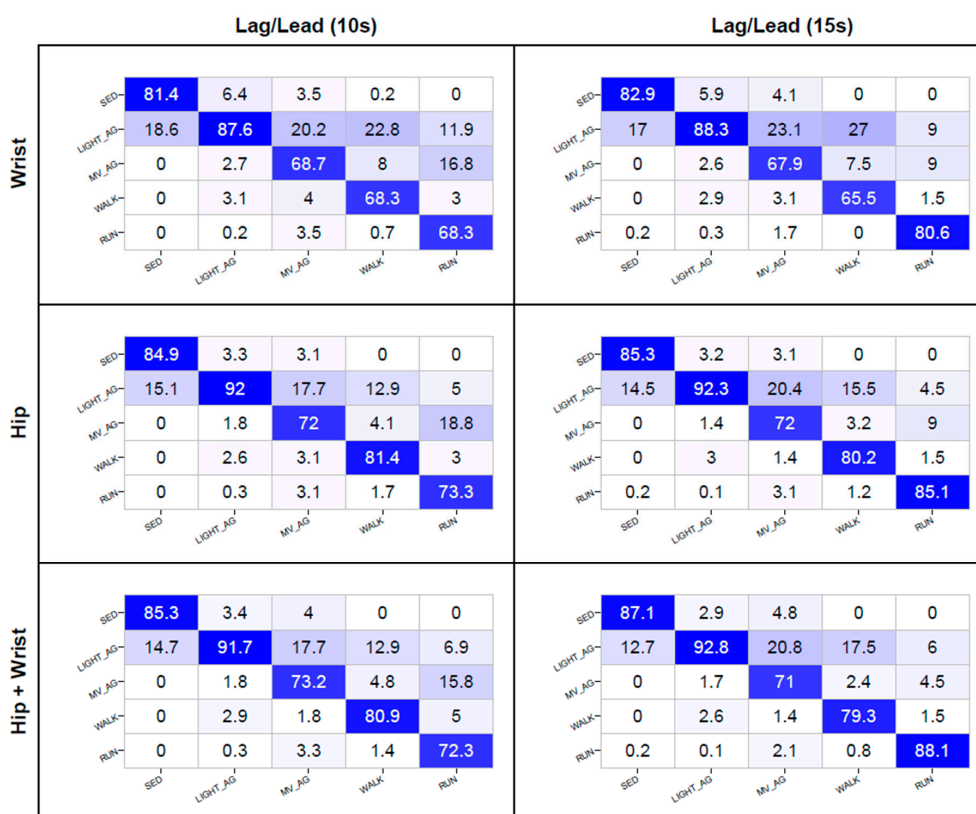
Placement	Feature	Window	SED	LIGHT_AG	MV_AG	WALK	RUN	Ave F-Score
Wrist	Base	1	63.3	71.4	45.7	45.9	55.1	62.6
		5	69.3	76.5	60.9	60.7	68.5	70.8
		10	73.7	79.1	62.0	68.8	73.4	74.5
		15	78.2	81.2	62.1	70.5	82.4	77.3
	Lag/Lead	1	69.2	75.5	57.5	54.8	61.6	68.8
		5	78.5	80.1	66.9	60.3	68.8	75.5
		10	82.4	82.9	70.3	70.8	71.5	80.0
		15	83.3	83.7	70.7	69.0	82.4	80.6



Table 2. Cont.

Placement	Feature	Window	SED	LIGHT_AG	MV_AG	WALK	RUN	Ave F-Score
Hip	Base	1	73.3	76.0	61.0	55.7	63.1	70.6
		5	80.6	82.7	75.5	69.8	71.2	79.5
		10	82.3	85.7	77.6	80.7	74.4	83.1
		15	85.0	86.8	75.3	78.4	80.0	84.0
	Lag/Lead	1	80.0	80.4	65.2	62.9	67.2	75.8
		5	85.7	85.6	76.3	68.8	72.8	82.2
		10	87.6	87.9	76.4	81.0	73.6	85.3
		15	87.7	88.2	78.5	79.4	82.6	85.9
Hip & Wrist	Base	1	75.8	77.7	62.1	58.4	64.5	72.5
		5	81.1	82.8	74.3	70.1	71.7	79.6
		10	83.9	85.8	75.6	79.5	74.9	83.2
		15	85.5	86.4	77.1	78.1	86.6	84.3
	Lag/Lead	1	80.8	81.0	66.7	64.9	67.9	76.8
		5	86.0	85.6	74.5	69.5	72.5	82.1
		10	87.5	87.9	77.2	80.7	73.0	85.3
		15	88.7	88.4	78.0	79.8	86.8	86.4

Heat map confusion matrices for the wrist, hip, and combined hip and wrist classifiers are presented in Figure 2. To reduce the complexity, only the results for the lag/lead models trained on 10 and 15 s windows are reported. Detailed confusion matrices for all 24 classification models can be found in the Supplementary Files.



**Figure 2.** Confusion matrices for physical activity classification from the wrist, hip, and combined hip and wrist placement for lag/lead 10 and 15 s window models. The columns represent observed; rows represent predictions; bold represents correct predictions; SED = sedentary; LIGHT\_AG = light physical activity and games; MV\_AG = moderate to vigorous physical activity and games; WALK = walking; RUN = running.

For the wrist placement, the lag/lead 10 and 15 s window models' recognition of SED (81.4%–82.9%) and LIGHT\_AG (87.6%–88.3%) were good. The SED instances were most frequently misclassified as LIGHT\_AG (17.0%–18.6%), while a small percentage of LIGHT\_AG instances were misclassified as either SED (5.9%–6.4%), MV\_AG (2.6%–2.7%), or WALK (2.9%–3.1%). The recognition of RUN was also good for the 15 s model (80.6%), but only modest for the 10 s model (68.3%). For the 15 s model, RUN was most frequently misclassified as either LIGHT\_AG (9.0%–11.9%) or MV\_AG (9.0%–16.8%), with only a very small percentage of the instances misclassified as WALK (1.5%–3.0%). The recognition of MV\_AG (67.9%–68.7%) and WALK (65.5%–68.3%) was modest, with approximately 20.0% to 27.0% of MV\_AG and WALK instances misclassified as LIGHT\_AG.

For the hip placement, the lag/lead 10 and 15 s window models' recognitions were good to excellent for SED and LIGHT\_AG (84.9% to 92.3%). The majority of SED instances were misclassified as LIGHT\_AG (14.5%–15.1%), while a small percentage of LIGHT\_AG was misclassified as either SED (3.2%–3.3%), WALK (2.6%–3.0%), or MV\_AG (1.4%–1.8%). The recognition of RUN was good for the 15 s window models (85.1%) and acceptable for the 10 s windows (73.3%). RUN was most frequently misclassified as MV\_AG (9.0%–18.8%). The recognition of WALK was good for the 10 and 15 s windows (80.2%–81.4%). WALK was most frequently misclassified as LIGHT\_AG (12.9%–15.5%). The recognition of MV\_AG was acceptable (72.0%), with 17.7% to 20.4% of instances misclassified as LIGHT\_AG.

For the combined hip and wrist classifiers, the 10 s window, and the 15 s window models, the recognition was good to excellent for SED and LIGHT\_AG (85.3%–92.8%). Almost all the misclassifications of SED instances were as LIGHT\_AG (12.7%–14.7%), and a small percentage of LIGHT\_AG was misclassified as either SED (2.9%–3.4%), WALK (2.6%–2.9%), or MV\_AG (1.7%–1.8%). The recognition of RUN was good for the 15 s window model (88.1%), and acceptable for the 10 s window model (72.3%). For the 10 s window model, RUN was most frequently misclassified as MV\_AG (15.8%), whilst for 15 s window model, misclassification occurred mostly as either LIGHT\_AG (4.5%) or MV\_AG (6.0%). The recognition of WALK was good for the 10 s and 15 s window models (79.3%–80.9%). WALK was most frequently misclassified as LIGHT\_AG (12.9%–17.5%). The recognition of MV\_AG was acceptable (71.0%–73.2%), with 17.7% to 20.8% of instances misclassified as LIGHT\_AG.

#### 4. Discussion

Machine learning approaches to accelerometer data processing have emerged as more versatile and potentially more accurate alternative to cut-point methods. However, when laboratory-trained activity classification models are evaluated under free-living conditions, substantial decreases in accuracy have been reported [23,25,33]. In the current study, the preschooler activity classification models trained on free-living accelerometer data exhibited a substantially higher accuracy than that reported for comparable models trained on laboratory data. Under free-living conditions, F-scores for the best performing wrist and hip model were 80.6% and 85.9%, respectively. In comparison, F-scores for the corresponding laboratory trained model was 60.2% for the wrist and 64.4% for the hip [26]. In contrast to the poor detection of walking reported for the laboratory-trained model, which ranged between just 8% and 15%, the free-living models correctly classified 68% to 82% of the walking instances.

Our findings are consistent with the results of adult studies comparing the performance of classification models trained with laboratory and free-living data under free-living conditions. Among older adults, Sasaki et al. [23] reported overall accuracies ranging from 49% to 55% for RF and support vector machine physical activity classification models trained on laboratory-based accelerometer data from the wrist, hip, or ankle. When the same models were trained on free-living data, the accuracy increased by 9–15 percentage points to 58% to 69%. Ermes et al. [33] trained decision tree and artificial neural network physical activity classification models using free-living data and observed an overall accuracy of 89% compared to 72% when the models were trained on laboratory data. Similarly, Bastian et al. [25] reported respective increases of 71 and 40 percentage points for the detection of sedentary activities and cycling after training a Bayesian activity classification model on both laboratory and



free-living data. These findings, together with the results of the current study, provide consistent evidence that machine learning activity classification models should be trained and tested with data collected under true free-living conditions.

The choice of window size or epoch length has significant implications for the accurate assessment of physical activity in young children [34–37]. Several studies have noted that young children perform physical activities intermittently in short bursts lasting a few seconds [38,39]. Therefore, for classification models predicting specific physical activity classes or types, it is important that the prediction windows be as short as possible so that the number of windows containing multiple activities can be minimized. In this study, it was hypothesized that a smaller window would reduce the proportion of “mixed windows” and therefore improve the classification accuracy. To test this hypothesis, the models were trained using window sizes of 1, 5, 10, and 15 s. Contrary to our hypothesis, the F-scores increased as the window size increased. While shorter prediction intervals reduced the number of windows with multiple activity types, the costs of mixed activity windows were more than offset by the benefits of having sufficient information to reliably capture movement patterns. Of the 24 models tested, only the wrist lag/lead model trained on 10 s data, and the hip and combined hip and wrist models with prediction windows of 10 or 15 s achieved average F-scores of  $\geq 80\%$  and provided F-scores  $\geq 70\%$  for all five activity classes.

For all the accelerometer placements, the models trained on 1 s windows did not provide an acceptable accuracy. The inferior performance of these models was due, in large part, to the poor recognition of MV\_AG, WALK, and RUN. Between 26% and 52% of the MV\_AG and WALK instances were misclassified as LIGHT\_AG, and between 24% and 39% of the RUN instances were misclassified as MV\_AG. It was evident that a prediction window of only 1 s provided insufficient information to reliably differentiate between these activity classes. With just 1 s of data (100 instances), features that differentiated MV\_AG and WALK from LIGHT\_AG (e.g., log energy, dominant frequency magnitude, entropy) over longer prediction windows were similar in magnitude; likewise, features that differentiated RUN from MV\_AG (e.g., mean absolute deviation, signal power, entropy) were similar in magnitude. With longer prediction windows, more data was accumulated, thereby increasing the precision with which discriminative features were calculated. Consequently, the detection accuracy increased for these activity classes.

In support of our hypothesis, the addition of features from lag and lead windows significantly improved the free-living classification accuracy. Temporal features have previously been used in studies developing and testing physical activity classification models, although their relative importance was not systematically evaluated. Zhang et al. [40] used the ratio of the dominant frequency recorded for the current and previous window to train decision tree and support vector machine activity classification models for the wrist and hip. For both accelerometer placements, the overall classification accuracy exceeded 95%, although the model was trained and tested using laboratory-based activity trials. Among preschool-aged children, Zhao et al. [41] used the lag and lead ActiGraph activity counts and step counts to train a support vector machine activity classifier. The overall accuracy exceeded 85%.

In the current study, the inclusion of temporal features based on the standard deviation of the signal vector magnitude was most beneficial for the models trained on the wrist accelerometer data. For these models, the inclusion of temporal features increased the overall F-score by an average of six percentage points, largely through the improved detection of SED, LIGHT\_AG, and MV\_AG. When performing physical activities in these three class categories, the accelerometer data recorded at the wrist exhibited significantly greater window-to-window variability than the accelerometer data recorded at the hip. Thus, for models trained on the wrist accelerometer data, the inclusion of temporal features related to signal variability in lag and lead windows was informative for classifying physical activities in these classes. Conversely, for clearly defined rhythmic activities such as walking and running, the temporal stability of the accelerometer signal was less dependent on the accelerometer placement.

Contrary to our hypothesis, the development of models based on multiple accelerometer placements and feature fusion did not improve the free-living classification performance. This could

be a function of the simple feature fusion approach implemented in the current study. Feature fusion approaches that use a class-based decision fusion technique may better capitalize on the increased information and features resulting from multiple accelerometer placements [42]. Alternatively, the lack of improvement might be due to a less than optimal combination of accelerometer placements. Narayanan et al. [43] reported a greater overall accuracy for a physical activity classifier trained on accelerometer data from the thigh and lower back than classifiers trained on data from the wrist and back and the wrist and thigh.

When evaluated under laboratory conditions, activity classification models trained on hip and wrist accelerometer data exhibit comparable accuracy [22,44,45]. However, in the current study, the classifiers trained on the hip accelerometer data exhibited significantly higher F-scores than the wrist models. The discrepancy in findings may be attributable, in part, to differences in how the machine learning classification models are trained and tested in laboratory-based studies compared to free-living studies. When classification models are trained in the laboratory setting, the participants complete a small number of choreographed activity trials which do not replicate how activities are performed in free-living scenarios. Moreover, when models are trained on free-living accelerometer data, the participants complete entirely different activities or complete the same activities in an entirely different manner. Furthermore, laboratory-based studies typically do not include activity trials that test the limitations of wrist mounted accelerometers. Sedentary and light intensity activities tend to be performed with limited upper limb movements (e.g., sitting on the floor listening to story, standing, and painting), while walking and running trials are performed continuously with arms oscillating freely. Under true free-living conditions, the participants often walk or run intermittently while holding objects, resulting in limited arm movements. This presents unique challenges for wrist classifiers and could account for the lower classification accuracy observed for the wrist models under true free-living conditions.

The free-living activity classifiers developed and validated in the current study can be implemented in field-based studies involving preschool-aged children and offer several advantages over traditional cut-point methods. By providing physical activity metrics based on the combination of activity type and intensity, the classification models allow researchers to monitor not only the quantity of physical activity, but also the quality of movement behaviors. For example, the classifiers can be used to measure how time in active game play is accumulated at home or early childhood education and care settings. Further, the study provides sufficiently accurate activity classification models for wrist mounted accelerometers. Wrist-worn accelerometers provide a higher compliance than accelerometers worn on the hip and allows researchers to concurrently measure sleep and evaluate adherence to newly released 24-hour movement guidelines. Furthermore, by classifying the activity class categories rather than the physical activity intensity based on energy expenditure, the models provide investigators with a measurement tool to examine age-related trends in movement behavior that are not confounded by developmental differences in the energy cost of physical activity [46–48].

The current study had several strengths. It is the first study to develop and evaluate the accuracy of machine learning activity classification models for preschool-aged children trained on free-living data. Ground-truth activity was obtained using a rigorous video-based direct observation and a continuous coding procedure. In addition, the study evaluated accuracy across multiple window sizes, feature sets, and accelerometer placements. Furthermore, mixed windows, in which multiple activities were performed within a given window, were not excluded from the evaluation which enhances the ecological validity and generalizability of the findings.

Opposing these strengths were several limitations. First, due to the demands of the data collection protocol, the free-living play sessions were restricted to 20 min. Although the participants were free to engage in any activity they chose at any location, longer play sessions may have provided a greater variety of moderate-to-vigorous intensity activities and more prolonged bouts of walking and running instances to train the models. Second, the study had a modest sample size of 31 participants. Nevertheless, the free play sessions generated more than 3.8 million data points and provided between

2500 and 38,500 activity windows, which was sufficient to train and test the models. Third, for classification models achieving average F-scores  $\geq 80\%$ , between 13% and 30% of the observed MV\_AG and WALK instances were misclassified as LIGHT\_AG. The propensity for activity classifiers trained solely on triaxial accelerometer data to misclassify certain activities in young children such as slow intermittent walking, walking while holding objects, and climbing on fixed playground equipment has been noted previously [26]. Therefore, to improve the recognition of MV\_AG and WALK, future studies should include features from additional sensors, such as heart rate monitors, gyroscopes, barometric pressure sensors, and GPS, which could provide additional information about physical activity intensity, posture, and changes in position and elevation. Fourth and finally, our study only trained RF classifiers and did not benchmark performance with other supervised learning algorithms. Because the focus of the current study was to evaluate activity classification models trained on true free-living data and compare the performance with models trained on laboratory data, it was necessary to employ the same supervised learning algorithm. Future studies could train and test classifiers using other machine learning algorithms and benchmark performance against the RF classifiers developed in the current study. The fully annotated training data for all window sizes are available at the following link: [https://github.com/QUTcparg/PS\\_PAClassification](https://github.com/QUTcparg/PS_PAClassification).

## 5. Conclusions

In summary, machine learning activity classification models trained on free-living data for preschool-aged children exhibited an acceptable accuracy under free-living conditions. The random forest activity classifiers with prediction windows of 10 or 15 s provided the accurate recognition of five activity classes representative of young children's movement behaviors. The inclusion of lag and lead features improved classification accuracy, with the largest increases observed for the wrist placement. The hip and combined hip and wrist classification models provided comparable accuracy, with both models outperforming the models trained on wrist accelerometer data. Future studies should train models using accelerometer data collected over extended time periods and a wider range of settings to provide more movement diversity in the training data. Such studies should explore the inclusion of additional temporal features, such as the ratio of the dominant frequency for the current and adjacent windows or information/features from additional sensors such as heart rate monitors, gyroscopes, barometric pressure sensors, and GPS trackers.

**Supplementary Materials:** The annotated training datasets, final RF models, and sample data with r scripts to run the classification models are available at the following link: [https://github.com/QUTcparg/PS\\_PAClassification](https://github.com/QUTcparg/PS_PAClassification).

**Author Contributions:** Conceptualization, M.N.A. and S.G.T.; methodology, M.N.A. and S.G.T.; formal analysis, M.N.A. and S.G.T.; writing—original draft preparation, M.N.A.; writing—review and editing, S.G.T. and T.G.P.; supervision, S.G.T. and T.G.P.; funding acquisition, S.G.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by AUSTRALIAN RESEARCH COUNCIL, grant number DP150100116.

**Acknowledgments:** We wish to thank the families for their participation and Denise K. Brookes for her contributions throughout the study.

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

## References

1. World Health Organization. *Report of the Commission on Ending Childhood Obesity*; World Health Organization: Geneva, Switzerland, 2016. [[CrossRef](#)]
2. De Onis, M.; Blossner, M.; Borghi, E. Global prevalence and trends of overweight and obesity among preschool children. *Am. J. Clin. Nutr.* **2010**, *92*, 1257–1264. [[CrossRef](#)] [[PubMed](#)]

3. Steinberger, J.; Daniels, S.R.; Hagberg, N.; Isasi, C.R.; Kelly, A.S.; Lloyd-Jones, D.; Pate, R.R.; Pratt, C.; Shay, C.M.; Towbin, J.A.; et al. Cardiovascular Health Promotion in Children: Challenges and Opportunities for 2020 and Beyond: A Scientific Statement From the American Heart Association. *Circulation* **2016**, *134*, e236–e255. [[CrossRef](#)] [[PubMed](#)]
4. Sonntag, D. Why Early Prevention of Childhood Obesity Is More Than a Medical Concern: A Health Economic Approach. *Ann. Nutr. Metab.* **2017**, *70*, 175–178. [[CrossRef](#)]
5. World Health Organization. *Prevalence of Insufficient Physical Activity*; Global Health Observatory; World Health Organization: Geneva, Switzerland, 2016; pp. 4–5.
6. Reinehr, T.; Hetherington, M.; Nekitsing, C.; Janet, M.; Shloim, N.; Ekelund, U. Aetiology of obesity in children. *Adv. Nutr. Diet. Obes.* **2018**, *169*, 261–286. [[CrossRef](#)]
7. Okely, A.D.; Ghersi, D.; Hesketh, K.D.; Santos, R.; Loughran, S.P.; Cliff, D.P.; Shilton, T.; Grant, D.; Jones, R.A.; Stanley, R.M.; et al. A collaborative approach to adopting/adapting guidelines—The Australian 24-Hour Movement Guidelines for the early years (Birth to 5 years): An integration of physical activity, sedentary behavior, and sleep. *BMC Public Health* **2017**, *17*, 869. [[CrossRef](#)] [[PubMed](#)]
8. Willumsen, J.; Bull, F. Development of WHO Guidelines on Physical Activity, Sedentary Behavior, and Sleep for Children Less Than 5 Years of Age. *J. Phys. Act. Health* **2020**, *17*, 96–100. [[CrossRef](#)] [[PubMed](#)]
9. Tremblay, M.S.; Chaput, J.-P.; Adamo, K.B.; Aubert, S.; Barnes, J.D.; Choquette, L.; Duggan, M.; Faulkner, G.; Goldfield, G.S.; Gray, C.E.; et al. Canadian 24-Hour Movement Guidelines for the Early Years (0–4 years): An Integration of Physical Activity, Sedentary Behaviour, and Sleep. *BMC Public Health* **2017**, *17*, 874. [[CrossRef](#)]
10. Cliff, D.; Reilly, J.J.; Okely, A.D. Methodological considerations in using accelerometers to assess habitual physical activity in children aged 0–5 years. *J. Sci. Med. Sport* **2009**, *12*, 557–567. [[CrossRef](#)]
11. Jago, R.; Zakeri, I.; Baranowski, T.; Watson, K. Decision boundaries and receiver operating characteristic curves: New methods for determining accelerometer cutpoints. *J. Sports Sci.* **2007**, *25*, 937–944. [[CrossRef](#)]
12. Welk, G.J. Principles of Design and Analyses for the Calibration of Accelerometry-Based Activity Monitors. *Med. Sci. Sports Exerc.* **2005**, *37*, 501–511. [[CrossRef](#)]
13. Trost, S.G.; Loprinzi, P.D.; Moore, R.; Pfeiffer, K.A. Comparison of Accelerometer Cut Points for Predicting Activity Intensity in Youth. *Med. Sci. Sports Exerc.* **2011**, *43*, 1360–1368. [[CrossRef](#)] [[PubMed](#)]
14. Hislop, J.F.; Bulley, C.; Mercer, T.H.; Reilly, J.J. Comparison of accelerometry cut points for physical activity and sedentary behavior in preschool children: A validation study. *Pediatr. Exerc. Sci.* **2012**, *24*, 563–576. [[CrossRef](#)] [[PubMed](#)]
15. Okely, A.D.; Batterham, M.J.; Hinkley, T.; Ekelund, U.; Brage, S.; Reilly, J.J.; Trost, S.G.; Jones, R.A.; Janssen, X.; Cliff, D.; et al. Wrist Acceleration Cut Points for Moderate-to-Vigorous Physical Activity in Youth. *Med. Sci. Sports Exerc.* **2018**, *50*, 609–616. [[CrossRef](#)] [[PubMed](#)]
16. Janssen, X.; Cliff, D.P.; Reilly, J.J.; Hinkley, T.; Jones, R.A.; Batterham, M.; Ekelund, U.; Brage, S.; Okely, A.D. Predictive Validity and Classification Accuracy of ActiGraph Energy Expenditure Equations and Cut-Points in Young Children. *PLoS ONE* **2013**, *8*, e79124. [[CrossRef](#)]
17. Kotsiantis, S.B. Supervised machine learning: A review of classification techniques. *Informatics* **2007**, *31*, 249–268.
18. Wang, H.; Ma, C.; Zhou, L. A Brief Review of Machine Learning and Its Application. In Proceedings of the 2009 International Conference on Information Engineering and Computer Science, Wuhan, China, 19–20 December 2009; pp. 1–4.
19. Mjolsness, E.; Decoste, D. Machine Learning for Science: State of the Art and Future Prospects. *Science* **2001**, *293*, 2051–2055. [[CrossRef](#)]
20. Trost, S.G.; Wong, W.-K.; Pfeiffer, K.A.; Zheng, Y. Artificial Neural Networks to Predict Activity Type and Energy Expenditure in Youth. *Med. Sci. Sports Exerc.* **2012**, *44*, 1801–1809. [[CrossRef](#)]
21. Ellis, K.; Kerr, J.; Godbole, S.; Staudenmayer, J.; Lanckriet, G. Hip and Wrist Accelerometer Algorithms for Free-Living Behavior Classification. *Med. Sci. Sports Exerc.* **2016**, *48*, 933–940. [[CrossRef](#)]
22. Trost, S.G.; Cliff, D.; Ahmadi, M.N.; Van Tuc, N.; Hagenbuchner, M. Sensor-enabled Activity Class Recognition in Preschoolers. *Med. Sci. Sports Exerc.* **2018**, *50*, 634–641. [[CrossRef](#)]
23. Sasaki, J.E.; Hickey, A.M.; Staudenmayer, J.W.; John, D.; Kent, J.A.; Freedson, P.S. Performance of Activity Classification Algorithms in Free-Living Older Adults. *Med. Sci. Sports Exerc.* **2016**, *48*, 941–950. [[CrossRef](#)]
24. Lyden, K.; Keadle, S.K.; Staudenmayer, J.; Freedson, P.S. A Method to Estimate Free-Living Active and Sedentary Behavior from an Accelerometer. *Med. Sci. Sports Exerc.* **2014**, *46*, 386–397. [[CrossRef](#)] [[PubMed](#)]



25. Bastian, T.; Maire, A.; Dugas, J.; Ataya, A.; Villars, C.; Gris, F.; Perrin, E.; Caritu, Y.; Doron, M.; Blanc, S.; et al. Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: Laboratory-based calibrations are not enough. *J. Appl. Physiol.* **2015**, *118*, 716–722. [[CrossRef](#)]
26. Ahmadi, M.N.; Brookes, D.; Chowdhury, A.; Pavey, T.; Trost, S.G. Free-living Evaluation of Laboratory-based Activity Classifiers in Preschoolers. *Med. Sci. Sports Exerc.* **2019**, *52*, 1227–1234. [[CrossRef](#)] [[PubMed](#)]
27. Koch, J.R.L.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159. [[CrossRef](#)]
28. Ahmadi, M.N.; Pfeiffer, K.A.; Trost, S. Physical Activity Classification in Youth Using Raw Accelerometer Data from the Hip. *Meas. Phys. Educ. Exerc. Sci.* **2020**, *24*, 1–7. [[CrossRef](#)]
29. Chowdhury, A.K.; Tjondronegoro, D.; Chandran, V.; Trost, S.G. Ensemble Methods for Classification of Physical Activities from Wrist Accelerometry. *Med. Sci. Sports Exerc.* **2017**, *49*, 1965–1973. [[CrossRef](#)]
30. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
31. Reiss, A.; Weber, M.; Stricker, D. Exploring and extending the boundaries of physical activity recognition. In Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics, Anchorage, AK, USA, 9–12 October 2011; pp. 46–50.
32. Forman, G.; Scholz, M. Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement. *ACM SIGKDD Explor. Newsl.* **2010**, *12*, 49–57. [[CrossRef](#)]
33. Ermes, M.; Parkka, J.; Mantjarvi, 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](#)]
34. Vale, S.; Santos, R.; Silva, P.; Soares-Miranda, L.; Mota, J. Preschool children physical activity measurement: Importance of epoch length choice. *Pediatr. Exerc. Sci.* **2009**, *21*, 413–420. [[CrossRef](#)]
35. McClain, J.J.; Abraham, T.L.; Brusseau, T.A.; Tudor-Locke, C. Epoch Length and Accelerometer Outputs in Children: Comparison to direct observation. *Med. Sci. Sports Exerc.* **2008**, *40*, 2080–2087. [[CrossRef](#)] [[PubMed](#)]
36. Trost, S.G.; McIver, K.L.; Pate, R.R. Conducting Accelerometer-Based Activity Assessments in Field-Based Research. *Med. Sci. Sports Exerc.* **2005**, *37*, S531–S543. [[CrossRef](#)] [[PubMed](#)]
37. Hislop, J.F.; Bulley, C.; Mercer, T.H.; Reilly, J.J. Comparison of epoch and uniaxial versus triaxial accelerometers in the measurement of physical activity in preschool children: A validation study. *Pediatr. Exerc. Sci.* **2012**, *24*, 450–460. [[CrossRef](#)] [[PubMed](#)]
38. Bailey, R.C.; Olson, J.; Pepper, S.L.; Porszasz, J.; Barstow, T.J.; Cooper, D.M. The level and tempo of children's physical activities: An observational study. *Med. Sci. Sports Exerc.* **1995**, *27*, 1033–1041. [[CrossRef](#)]
39. Baquet, G.; Stratton, G.; Van Praagh, E.; Berthoin, S. Improving physical activity assessment in prepubertal children with high-frequency accelerometry monitoring: A methodological issue. *Prev. Med.* **2007**, *44*, 143–147. [[CrossRef](#)]
40. Zhang, S.; Rowlands, A.V.; Murray, P.; Hurst, T.L. Physical Activity Classification Using the GENE A Wrist-Worn Accelerometer. *Med. Sci. Sports Exerc.* **2012**, *44*, 742–748. [[CrossRef](#)]
41. Zhao, W.; Adolph, A.L.; Puyau, M.R.; Vohra, F.A.; Butte, N.F.; Zakeri, I.F. Support vector machines classifiers of physical activities in preschoolers. *Physiol. Rep.* **2013**, *1*, e0006. [[CrossRef](#)]
42. Chowdhury, A.K.; Tjondronegoro, D.; Chandran, V.; Trost, S.G. Physical Activity Recognition Using Posterior-Adapted Class-Based Fusion of Multiaccelerometer Data. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 678–685. [[CrossRef](#)]
43. Narayanan, A.; Stewart, T.; Mackay, L. A Dual-Accelerometer System for Detecting Human Movement in a Free-living Environment. *Med. Sci. Sports Exerc.* **2020**, *52*, 252–258. [[CrossRef](#)]
44. Trost, S.G.; Zheng, Y.; Wong, W.-K. Machine learning for activity recognition: Hip versus wrist data. *Physiol. Meas.* **2014**, *35*, 2183–2189. [[CrossRef](#)]
45. Cleland, I.; Kikhia, B.; Nugent, C.; Boytsov, A.; Hallberg, J.; Synnes, K.; McClean, S.; Finlay, D.D. Optimal Placement of Accelerometers for the Detection of Everyday Activities. *Sensors* **2013**, *13*, 9183–9200. [[CrossRef](#)] [[PubMed](#)]
46. Trost, S.G.; Drovandi, C.; Pfeiffer, K. Developmental Trends in the Energy Cost of Physical Activities Performed by Youth. *J. Phys. Act. Health* **2016**, *13*, S35–S40. [[CrossRef](#)] [[PubMed](#)]

47. Harrell, J.S.; McMurray, R.G.; Baggett, C.D.; Pennell, M.L.; Pearce, P.F.; Bangdiwala, S.I. Energy Costs of Physical Activities in Children and Adolescents. *Med. Sci. Sports Exerc.* **2005**, *37*, 329–336. [[CrossRef](#)]
48. McMurray, R.G.; Butte, N.F.; Crouter, S.E.; Trost, S.G.; Pfeiffer, K.A.; Bassett, D.R.; Puyau, M.R.; Berrigan, D.; Watson, K.B.; Fulton, J.E.; et al. Exploring Metrics to Express Energy Expenditure of Physical Activity in Youth. *PLoS ONE* **2015**, *10*, e0130869. [[CrossRef](#)] [[PubMed](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).