Evaluating and Enhancing the Generalization Performance of Machine Learning Models for Physical Activity Intensity Prediction From Raw Acceleration Data

Vahid Farrahi, Maisa Niemelä, Petra Tjurin, Maarit Kangas, Raija Korpelainen, and Timo Jämsä, *Senior Member, IEEE*

***Abstract*—Purpose: To evaluate and enhance the generalization performance of machine learning physical activity intensity prediction models developed with raw acceleration data on populations monitored by different activity monitors. Method: Five datasets from four studies, each containing only hip- or wrist-based raw acceleration data (two hip- and three wrist-based) were extracted. The five datasets were then used to develop and validate artificial neural networks (ANN) in three setups to classify activity intensity categories (sedentary behavior, light, and moderate-to-vigorous). To examine generalizability, the ANN models were developed using within dataset (leave-one-subject-out) cross validation, and then cross tested to other datasets with different accelerometers. To enhance the models' generalizability, a combination of four of the five datasets was used for training and the fifth dataset for validation. Finally, all the five datasets were merged to develop a single model that is generalizable across the datasets (50% of the subjects from each dataset for training, the remaining for validation). Results: The datasets showed high performance in within dataset cross validation (accuracy 71.9–95.4%, Kappa K = 0.63–0.94). The performance of the within dataset validated models decreased when applied to datasets with different accelerometers (41.2–59.9%, K = 0.21–0.48). The trained models on merged datasets consisting hip and wrist data predicted the left-out dataset with acceptable performance (65.9–83.7%, K = 0.61–0.79). The model trained with all five datasets performed with acceptable performance across the datasets (80.4–90.7%, K = 0.68–0.89). Conclusions: Integrating heterogeneous datasets in training sets seems a viable approach for enhancing the generalization performance of the models. Instead, within dataset validation is not sufficient to understand the models' performance on other populations with different accelerometers.**

***Index Terms*—Accelerometers, pattern recognition, artificial neural networks, activity monitor, classification.**

Manuscript received January 7, 2019; revised March 21, 2019 and May 1, 2019; accepted May 14, 2019. Date of publication May 20, 2019; date of current version January 6, 2020. This work was supported in part by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie under Grant 713645, in part by the Ministry of Education and Culture in Finland under Grants OKM/86/626/2014, OKM/43/626/2015, OKM/17/626/2016, and OKM/54/626/2019, and in part by Infotech Oulu, Finland, and Northern Ostrobothnia Hospital District. *(Corresponding author: Vahid Farrahi.)*

V. Farrahi is with the Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu 90014, Finland (e-mail: vahid.farrahi@oulu.fi).

M. Niemelä is with the Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu 90014, Finland, and with the Infotech, University of Oulu, Oulu 90014, Finland, and also with the Medical Research Center, Oulu University Hospital and University of Oulu, Oulu 90014, Finland (e-mail: maisa.niemela@oulu.fi).

P. Tjurin is with the Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu 90014, Finland, and also with the Oulu Deaconess Institute Foundation, Department of Sports and Exercise Medicine, Oulu 90100, Finland (e-mail: petra.tjurin @oulu.fi).

M. Kangas is with the Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu 90014, Finland, and also with the Medical Research Center, Oulu University Hospital and University of Oulu, Oulu 90014, Finland (e-mail: maarit.kangas@oulu.fi).

R. Korpelainen is with the Medical Research Center, Oulu University Hospital and University of Oulu, Oulu 90014, Finland, and with the Center for Life Course Health Research, University of Oulu, Oulu 90014, Finland, and also with the Oulu Deaconess Institute Foundation, Department of Sports and Exercise Medicine, Oulu 90100, Finland (e-mail: raija.korpelainen@odl.fi).

T. Jämsä is with the Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu 90014, Finland, and with the Medical Research Center, Oulu University Hospital and University of Oulu, Oulu 90014, Finland, and also with the Department of Diagnostic Radiology,Oulu University Hospital, Oulu 90220, Finland (e-mail: timo.jamsa@oulu.fi).

Digital Object Identifier 10.1109/JBHI.2019.2917565

I. Introduction

ACCELEROMETERS are small, reliable, and feasible tools for objective measurement of physical activity (PA) [1]. Different PA types, energy expenditure, and intensities can be assessed from acceleration signals [1]. Classification of acceleration data across the whole intensity spectrum is one of the most common measures for a variety of studies including clinical, surveillance, and intervention studies [2].

Traditionally, intensity classification of activities has been performed using cut-point-based methods that have been established for both activity counts and raw acceleration data [1], [2]. However, the accuracy of the cut-points has been reported to be limited [1], [3]. Recently, raw accelerometry and machine learning (ML) modeling approaches have been used for both standardization and harmonization of accelerometry results and precise measurement [4], [5]. It is widely accepted that the increased output comparability provided by raw accelerometry, together with sophisticated modeling approaches, could help develop reliable and precise data processing techniques capable of predicting PA in various population groups regardless of accelerometer brand [6]–[8]. Such a technique is needed to enable the comparability of results across studies and provide opportunities to pool data from different cohorts [9]. To date, a universally accepted method for predicting PA intensity from acceleration data across the intensity spectrum is still lacking [2], [8]. This is mainly because the generalizability of the existing methods to populations different from the one used for model development has been limited. Several factors affecting the generalization performance of modeling approaches. Studies focusing on the comparability of accelerometry data have found that raw data from various accelerometers are not equivalent [6], [7], resulting in incomparability of activity intensity predictions [6]. Differences in population characteristics also affect the generalization performance of the existing methods. These led to requests for more robust data processing techniques capable of predicting PA intensities in various populations monitored by different accelerometers [6], [7]. Recent evidence has suggested developing ML-based classification models to classify activities directly according to their intensity categories. The validity of this method [10], [11] as well as its generalizability on independent population [12] have been shown in previous studies. This method might be preferred over indirect methods, which first predict the energy expenditure of activities and then classify activity intensity based on energy expenditure thresholds [13], or first predict activity types and then collapse the categories into intensity categories [11], [12]. Previous research has reported both high error and bias of energy expenditure prediction models [12], [13] and performance deterioration of classification models as the number of activity categories increases [14], [15].

The present study focuses on evaluating and enhancing generalization performance on independent populations monitored by different accelerometers. The specific study goals were (a) to examine the generalization performance of PA intensity prediction models developed with raw acceleration data and validated using within-sample validation to other populations monitored with different accelerometers, (b) to investigate whether the generalization performance of intensity prediction models can be improved by training the models on data collected from different populations with different wear locations (hip or wrist) and accelerometers, and (c) to provide a robust intensity prediction model that is trained on a variety of data sources and performs across different populations monitored by different accelerometers and wear locations.

The article is organized as follows. The related works are described briefly in Section II. The dataset and data preparation steps are presented in Section III, followed by the experimental design in Section IV. The results are shown in Section V and discussed in detail in Section VI. Finally, the conclusions of the study are presented in Section VII.

II. Related Works

A. Machine Learning Approaches for Physical Activity Intensity Prediction

Pober *et al.* were one of the first to develop and validate ML-based models for classifying PA from acceleration data [16]. They also demonstrated that ML-based models are more precise for predicting PA intensities compared with regression analysis using PA intensity cut-points. With the accumulating evidence on the superiority of ML approaches, questions have been raised about which technique is preferred. This has led to the examination of different ML techniques for predicting PA types from different wear locations in existing studies. For example, Zhang *et al.* [17] tested several ML techniques with a wrist-worn accelerometer and achieved the highest accuracy by the support vector machine and decision tree, while other tested ML methods including artificial neural networks (ANN), Naïve Bayes, and logistic regression also produced satisfactory results. Another study by Kate *et al.* [18], testing different ML techniques with a hip-worn accelerometer, achieved the highest accuracy using ANN, logistic regression, and support vector machines compared to random forest, Naïve Bayes, decision tree, bagged decision tree, and K-nearest neighbors. To date, there is still no consensus in the literature on which ML technique is the most accurate, mainly because their performance can depend on different factors such as wear location, population age range, window size, and even defined activity categories that can vary from one study to another [5], [12]. While various ML approaches are still being validated, recent evidence suggests the popularity of ANN compared to other applied approaches [4], [5]. This might be mainly due to the repeated documentation of its reasonably high and competitive predictive accuracy with different wear locations [14], populations [12], and window sizes [19] compared to other applied ML techniques. It has also been known as a robust and flexible approach for predicting PA types from raw acceleration data [4].

Another main concern regarding the use of ML approaches for predicting PA has been their generalization performance on independent populations especially for predicting PA intensity [1]. Previous studies have been mostly developed using within-sample validation (i.e., leave one subject out), resulting in a limited understanding of the performance of ML-based models on different populations [4], [5].

B. Previous Studies Testing the Generalization of Machine Learning-Based Models

To date, there has been only a few studies which investigated the generalization performance of ML-based models [5]. Although studies have consistently shown a performance deterioration when ML models are cross-tested on independent populations, there is still no consensus on its cause. However, the reason identified for this deterioration has been different among these studies [5]. For example, while Bastian *et al.* identified that the differences between acceleration data collected in free-living and laboratory settings is the main reason for

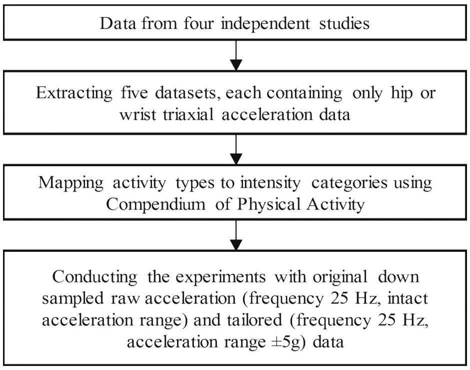
Data from four independent studies
Extracting five datasets, each containing only hip or wrist triaxial acceleration data
Mapping activity types to intensity categories using Compendium of Physical Activity
Conducting the experiments with original down sampled raw acceleration (frequency 25 Hz, intact acceleration range) and tailored (frequency 25 Hz, acceleration range ±5g) data

Fig. 1. The general workflow of study.

the performance degradation of laboratory-calibrated models when applied to an independent population [20], Mannini *et al.* identified the differences in sample characteristics as the main reason for degradation [21]. These discrepancies among previous studies seem to be due to the investigation of only one factor at a time whereas there would be several contributing factors in real-world applications, ranging from acceleration data to sample characteristics and unseen activities [5]. It is necessary to understand how ML-based models developed for PA intensity prediction will perform in the presence of various heterogeneities and to develop more robust models.

III. Datasets and Data Preparation

The present study conducted three experiments using data from a total of four independent studies comprising raw acceleration-based motion data. This section first describes these studies and how they collected motion data. Then, it describes how these datasets were prepared for the experiments. The study workflow is shown in Fig. 1.

A. Datasets

One of the four studies (University of Oulu) was performed by our research team at the University of Oulu [10] and the other three had open-access data (Oregon State University [22], PAMAP2 Physical Activity Monitoring [23], Daily and Sports Activities [24]). These studies were performed by different research labs with different participant groups (i.e., youth and adults), and each of which comprised different sets of activities performed in different contexts (indoor and outdoor) measured by different activity monitors. The open-access datasets were selected since, to the best of our knowledge, they were the only ones that were publicly available to the research community at the time of the study, and comprised raw acceleration data measured by wearable activity monitors. A brief description of each study is provided in Table I, and the performed activities are presented provided in Table II.

1) Dataset UOULU (Hip): This dataset was gathered at the University of Oulu (UOULU). It was collected from 22 healthy adult participants who performed 10 types of pre-defined activities [10]. The participants were recruited from the students and staff members of the University of Oulu. Direct observation served as the criterion measure to annotate the data. The Hookie AM20 triaxial accelerometer was attached to the participant's right hip to measure raw acceleration data (scale: 16 gravitational unit [g]) with a sampling rate of 100 Hz. The data collection protocol included activities in the following order: lying on a sofa, working on a computer, standing, table cleaning, floor cleaning, walking slow (approximately 2–3 km/h), walking fast (approximately 5–6 km/h), playing soccer, jogging, and cycling. Each activity trial lasted for 4 min except laying on a sofa, which lasted for 5.5 min. In between each activity, participants rested for 1 to 6 min. Detailed information about the data acquisition protocol has been reported previously [10].

2) Dataset OSU (Hip, Wrist): The dataset has been made publicly available by researchers at Oregon State University (OSU) [25]. It was collected from 52 youths performing 12 types of activities in two lab visits within two weeks and annotated by direct observation. The Actigraph GT3X+ triaxial accelerometers were attached to participant's right hip and non-dominant wrist to measure raw acceleration data (scale: ±6 g) with a sampling rate of 30 Hz. Briefly, on the first lab visit, participants completed six of the activities, and on the second lab visit, they completed the remaining six activities. Each activity trial lasted 5 min. More information about the activity protocol can be found elsewhere [22]. The authors specified that the data were collected from 52 subjects. However, the hip data from 50 subjects and wrist data from 16 were available for download.

3) Dataset PAMAP (Wrist): The PAMAP2 Physical Activity Monitoring (PAMAP) dataset is a fully annotated dataset by direct observation. It is publicly available from the UCI Machine Learning Repository [26]. The dataset was acquired from 9 adult participants who performed 18 different types of activities. The subjects were mainly employees or students at the German Research Center of Artificial Intelligence. To capture motion data, three Colibri wireless inertial measurement units (IMU) were attached to participants' chest and dominant wrist and ankle. Each of the IMUs contained two triaxial acceleration sensors (scale: ±16 g and ±6 g); a 3-D gyroscope sensor; a 3-D magnetometer sensor; and temperature, orientation, and HR monitor sensors. The raw acceleration data from both accelerometers were acquired with 100 Hz sampling rate. Briefly, each participant followed a protocol containing 12 different activities. Additionally, participants were asked to perform some additional activities from a list. Five of the participants did the additional activities, including watching TV, driving a car, playing soccer, and folding laundry. Detailed information about the activity protocol can be found elsewhere [23].

4) Dataset DSA (Wrist): The Daily and Sports Activities (DSA) dataset is another publicly available dataset. The data and its annotation information are available from UCI Machine Learning Repository [27]. It was collected from 8 adult subjects performing 18 different types of activities. Five Xsens MTx IMUs were attached to the chest and the right and left wrists and knees to capture the motion data. Each of the IMUs contained a triaxial acceleration sensor, a 3-D gyroscope sensor, and a 3-D magnetometer sensor. The accelerometers of the IMUs attached to the wrists had a scale of ±5 g and the other three had a

TABLE I
DETAILED CHARACTERISTICS OF THE FOUR STUDIES

Dataset	Collection environment	Participants	Sensors and accelerometer sensor specifics	Wear locations
University of Oulu (UOULU)	A dataset collected inside and outside a laboratory.	Twenty two participants (11 males, 11 females), mean age: 27.5 ± 11.2 year, age range: 17-58 year, BMI: 25.1 ± 2.2 kg.m^{-2}	Hookie AM20 triaxial accelerometer, scale: ±16g, sampling rate: 100 Hz	Right hip
Oregon State University (OSU)	An open-access dataset collected inside a laboratory.	Fifty two participants (28 boys, 24 girls), mean age: 13.7 ± 3.1 year, age range: 7.2-18.9 year, body mass: 50.6 ± 13.5 kg, handedness: not specified	ActiGraph GT3X+ triaxial accelerometer; scale: ±6g, sampling rate: 30 Hz	Right hip and non-dominant wrist
PAMAP2 Physical Activity Monitoring (PAMAP),	An open-access dataset collected inside and outside a laboratory.	Nine participants (8 males, 1 females), mean age: 27.2 ± 3.3 year, age range: 23-32 year, BMI: 25.1 ± 2.6 kg.m^{-2}, handedness: 7 right, 1 left	Colibri wireless inertial measurement unit containing two 3D accelerometers, 3D gyroscope sensor, 3D magnetometer sensor, temperature, orientation and HR monitor sensors; scales: ±16g and ±6g, sampling rate: 100 Hz	Dominant-wrist, dominant-ankle, chest
Daily and Sports Activities (DSA)	An open-access dataset collected inside and outside a laboratory.	Nine participants (4 males, 4 females), mean age: not specified, age range: 20-30 year, BMI: not specified, handedness: not specified	Xsens MTx wired inertial measurement unit containing 3D accelerometer, 3D gyroscope, 3D magnetometer; scales: ±5g (wrists) and ±18g (knees and chest), sampling rate: 25 Hz	Right and left wrists, right and left knees, chest

scale of ± 18 g. The raw acceleration data were collected with a sampling rate of 25 Hz. Each of the 18 activities were performed by all 8 subjects for 5 min. Further information regarding the activity protocol is described elsewhere [24].

B. Dataset Preparation

Raw acceleration data from the four abovementioned studies were extracted and used to create five datasets, each containing only hip or wrist triaxial raw acceleration data. Throughout the text, the five datasets are referred to as UOULU (H), OSU (H), OSU (W), PAMAP (W), and DSA (W), where H refers to hip and W to wrist. The dataset PAMAP (W) includes the wrist acceleration data from the IMU sensor with a scale of ± 16 g. The dataset DSA (W) includes the right wrist acceleration data.

1) Mapping Activity Types to Intensity Categories: In all the datasets, direct observation served as the criterion for physical activity. The Compendium of Physical Activity for adults [28] and youths [29] were used to assess the energy expenditure associated with each activity in the adult (UOULU, PAMAP, DSA) and youth (OSU) dataset, respectively. The Compendium of Physical Activity is a widely accepted tool and has been used with direct observation in previous studies as criterion measures for defining activity intensities [11], [12]. Based on the anatomical postures suggested by the Sedentary Behavior Research Network (SBRN) [30] and absolute MET (metabolic equivalent) thresholds, the activities were categorized into three intensity categories: ≤ 1.5: sedentary behavior (SB), 1.6–2.9: light PA (LPA), and ≥ 3.0: moderate-to-vigorous PA (MVPA). The performed activities and their corresponding codes in the Compendium of Physical Activity and intensity categories are displayed in Table II. The donut charts in Fig. 2 show the percentage distribution of the three intensity categories for all the subjects in each dataset.

2) Sampling Frequency and Acceleration Range: The datasets had different sampling frequencies ranging from 25 Hz to 100 Hz. Previous studies have shown that sampling frequency can slightly affect the results of classification models. To minimize the effect of sampling frequency, acceleration data in all the five datasets were downsampled to 25 Hz, which is previously shown to be enough for activity classification [31]. Another factor which can possibly affect the results of predictions is acceleration range. To date, it remains unknown which acceleration range is enough for predicting activity intensity from raw acceleration data. Therefore, we performed all the experiments both with the downsampled intact original raw acceleration data and with tailored data. In the experiments with tailored acceleration, the acceleration range was limited to ± 5 g, which is the minimum acceleration range across the datasets.

IV. EXPERIMENTAL DESIGN, MODELING APPROACH, AND PERFORMANCE ANALYSIS

This section first covers how the experiments were designed to evaluate and enhance the generalization performance of ML models for PA intensity prediction. Then, it describes the modeling approach followed by the performance evaluation and statistical analysis.

A. Experiments

1) Experiment 1: Leave-one-subject-out (LOSO) cross-validation is a commonly used validation approach in existing studies [4]. In this approach, data from all but one participant are used to train the model, and the left-out participant's data is used to validate the model; the procedure is repeated until all the participants are tested [32]. This experiment was designed to clarify the generalization performance of models validated within one population to independent ones with different characteristics

TABLE II
PERFORMED ACTIVITIES IN THE DATASETS AND, THEIR CORRESPONDING CODES IN THE ADULT [28] OR YOUTH [29] COMPENDIUM OF PHYSICAL ACTIVITY, AND ACTIVITY INTENSITY CATEGORIES

Activity type	Compendium code	Activity intensity
*University of Oulu (UOULU)^*		
Lying on a sofa	07011	SB
Computer work	11580	SB
Standing	07040	LPA
Table cleaning	05042	LPA
Floor cleaning	05011	LPA
Walking slow 2-3 km/h	17151	LPA
Walking fast 5-6 km/h	17200	MVPA
Playing soccer	15610	MVPA
Jogging	12020	MVPA
cycling	01010	MVPA
*Oregon State University (OSU)**		
Lying	50100	SB
Hand writing	55500	SB
Computer game	35200	SB
Sweeping	45180	LPA
Underarm throw and catch	20100	LPA
Aerobic dancing	40100	MVPA
Laundry task	45240	MVPA
Self-paced walking	80120	MVPA
Brisk walking	80300	MVPA
Brisk treadmill walk	80240	MVPA
Playing basketball	65100	MVPA
Jogging	60140	MVPA
*PAMAP2 Physical Activity Monitoring (PAMAP)^*		
Lying	07011	SB
Sitting	07021	SB
Watching TV	07020	SB
Computer work	11580	SB
Standing	07040	LPA
Car driving	16010	LPA
Folding laundry	05090	LPA
House cleaning	05025	LPA
Walking	17190	MVPA
Running	12020	MVPA
Nordic walking	17302	MVPA
Ascending stairs	17133	MVPA
Descending stairs	17070	MVPA
Vacuum cleaning	05043	MVPA
Playing soccer	15610	MVPA
Rope jumping	15552	MVPA
Cycling	01010	MVPA
*Daily and Sports Activities (DSA)^*		
Lying on back	07011	SB
Lying on right side	07011	SB
Sitting	07021	SB
Standing	07040	LPA
Standing in an elevator	07040	LPA
Moving around in an elevator	17151	LPA
Walking in a parking lot	17151	LPA
Rowing	18040	LPA
Ascending stairs	17133	MVPA
Descending stairs	17070	MVPA
Walking on a treadmill 4 km/h	17170	MVPA
Walking on a treadmill 4 km/h in 15 degree inclined position	17211	MVPA
Running on a treadmill 8 km/h	12030	MVPA
Exercising on a stepper	03018	MVPA
Exercising on a cross trainer	02048	MVPA
Stationary cycling in horizontal position	02010	MVPA
Stationary cycling in vertical position	02010	MVPA
Jumping	15552	MVPA
Playing basketball	15055	MVPA

*Youth Compendium of Physical Activity was used. ^Adult Compendium of Physical Activity was used. SB: Sedentary behavior, LPA: Light physical activity, MVPA: Moderate-to-vigorous physical activity.

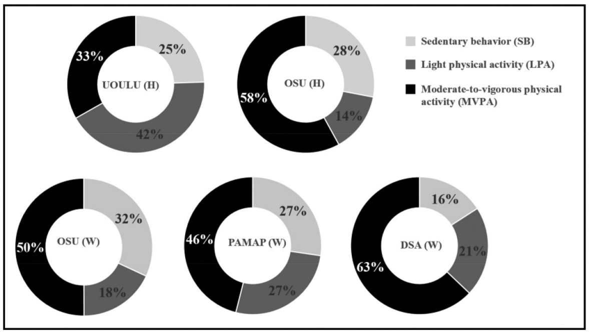

Fig. 2. Percentage distribution of the three intensity categories (SB, LPA, and MVPA) for all the subjects in the five datasets.

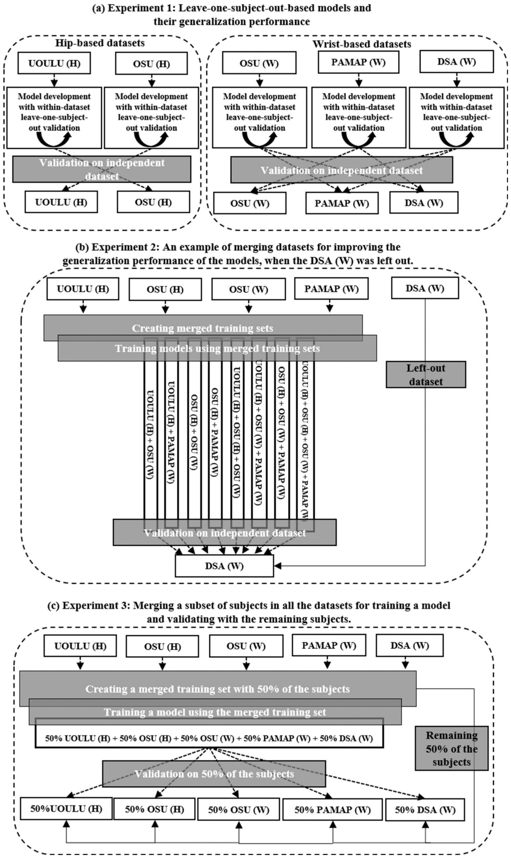
(a) Experiment 1: Leave-one-subject-out-based models and their generalization performance
Hip-based datasets
Wrist-based datasets
UOULU (H)
OSU (H)
OSU (W)
PAMAP (W)
DSA (W)
Model development with within-dataset leave-one-subject-out validation
Validation on independent dataset
(b) Experiment 2: An example of merging datasets for improving the generalization performance of the models, when the DSA (W) was left out.
Creating merged training sets
Training models using merged training sets
UOULU (H) + OSU (W)
UOULU (H) + PAMAP (W)
OSU (H) + OSU (W)
OSU (H) + PAMAP (W)
UOULU (H) + OSU (H) + OSU (W)
UOULU (H) + OSU (W) + PAMAP (W)
OSU (H) + OSU (W) + PAMAP (W)
UOULU (H) + OSU (H) + OSU (W) + PAMAP (W)
Left-out dataset
Validation on independent dataset
DSA (W)
(c) Experiment 3: Merging a subset of subjects in all the datasets for training a model and validating with the remaining subjects.
Creating a merged training set with 50% of the subjects
Training a model using the merged training set
50% UOULU (H) + 50% OSU (H) + 50% OSU (W) + 50% PAMAP (W) + 50% DSA (W)
Remaining 50% of the subjects
Validation on 50% of the subjects
50%UOULU (H)
50% OSU (H)
50% OSU (W)
50% PAMAP (W)
50% DSA (W)

Fig. 3. General schema of the three experiments.

and accelerometers performing different sets of activities from the one used to develop the model. To achieve the goal, the placement-specific classification models were validated using LOSO cross-validation within each dataset, and the most optimal fit with the highest accuracy for each dataset was obtained. The final models were trained with the data of all participants. The optimal established models were then cross-validated on independent populations with similar accelerometer placement (out-of-sample testing) to evaluate their generalization performance. Fig. 3(a) demonstrates how Experiment 1 was designed to test the generalization performance of LOSO-based validated models.

2) Experiment 2: This experiment was conducted to test whether the generalization performance of intensity prediction models on independent populations can be improved and more robust models can be provided by incorporating the information that acceleration data from different body sites (i.e., hip or wrist) acquired from various populations might contain. For this, to keep the validation set independent, one dataset was used for validation at a time and was left out from model development (it was not used as training data). Then, using the remaining datasets, a merged dataset consisting of a combination of both hip- and wrist-based datasets was built and used as a training set to train an intensity classification model. The trained model using the merged training set was then validated with the left-out population. To find the most optimal model, all possible dataset combinations were tested to find the optimal training sets achieving the highest accuracy in predicting activity intensity categories in the left out dataset. For each created merged training set, the most optimal fit was found, and the most optimal model for classifying the left out was selected as the final model. These steps were repeated when one of the five datasets was left out at a time to find the most optimal merged training set for all the five left-out datasets. As an example, Fig. 3(b) demonstrates how the datasets were merged when the DSA (W) was left out.

3) Experiment 3: This experiment was conducted to test whether it is possible to develop a single classification model that performs acceptably on different populations wearing a single hip- or wrist-worn accelerometer. To test this possibility, a merged training set comprising data from all the hip- and wrist-based datasets was created and used to train a single model for classifying PA intensities across the five datasets. For this, the data of 50% of the subjects from each dataset were selected randomly, and their data were merged in a dataset for training the model. The data of the remaining 50% of the subjects in each dataset were then used to validate how the model performs across all the datasets. Data splitting by subject may lead to better representative data subsets for model development and validation, compared with splitting the whole dataset into half. Previous studies have also used this method for validating the performance of ML models [20]. Fig. 3(c) depicts how a single model was trained using a subset of subjects selected from all datasets and validated with all the remaining subjects.

B. Modeling Approach

1) Feature Extraction: The same feature sets were extracted for all the five dataset/placements. The nonoverlapping 60-second window length was chosen to segment the data. This window length has been used in recent studies to establish activity type classification and energy expenditure estimation models for youths [19] and adults [15]. For each interval, time- and frequency-domain features (obtained using a fast Fourier transform) were extracted from all the three axes of measurement (i.e., x, y, and z) and also for the vector magnitude (i.e., $\sqrt{x^2 + y^2 + z^2}$). Time-domain features included the 10th, 25th, 50th, 75th, and 90th percentiles of acceleration signals. Frequency-domain features included the 10th, 25th, 50th, 75th, and 90th percentiles of signal frequency, total signal power, dominant frequency, and dominant frequency between 0.6 and 2.5 Hz. The extracted feature set is very similar to those that were used and validated in previous works [11], [33].

2) Classification Algorithm: The extracted features were used as inputs to develop ML models. Artificial neural networks (ANNs) were selected to classify the three different activity intensities: SB, LPA, and MVPA. ANN, composed of mathematical nonlinear functions, can model complex relationships between inputs and outputs [19]. ANN was chosen because it has been used frequently in previous studies, and it has been reported to be highly accurate in predicting both activity type and energy expenditure from accelerometer data in different age groups [11], [19]. Further description of the theoretical basis and structure of ANN can be found in past studies focusing on developing ANN models to predict activity types or intensities [11], [19].

In this study, the *nnet* package in R was used to train the ANNs. Each network comprised a single hidden layer with 10 nodes. To ensure that the models are converged, each network was trained for a maximum of 10000 iterations. These parameters were selected based on reasonably high accuracy in previous works developing activity and intensity prediction models [11], [12].

C. Performance Evaluation and Statistical Analysis

Confusion matrices, showing the proportion of instances of the three intensity categories that were correctly and incorrectly classified, were used to evaluate the classification and misclassification rate of the ANN models. An overall classification accuracy of the models with 95% confidence interval was also produced by calculating the percentage of correctly classified 60-second time windows for each subject. To ensure that the overall predication accuracy was not by chance, the agreement between predicted and actual intensity categories was calculated by weighted Kappa statistics (K), which is known as an appropriate performance metric for multiclass classification tasks [34]. We followed the common and accepted interpretation of Kappa values (k) suggested by Landis and Koch [35], where the following categorization is used: poor (0.0–0.2), fair (0.2–0.4), moderate (0.4–0.6), substantial (0.6–0.8), and almost perfect (0.8–1.0).

Besides classification accuracy, it is also important to test whether a single model is providing comparable results for different intensity categories across the dataset [9]. Thus, in Experiment 3, we further tested the comparability of classification results by calculating sensitivity and specificity for each intensity category across the datasets. The means of sensitivity and specificity were compared to test whether they differ significantly. One-way analysis of variance (ANOVA) with the Tukey-Kramer post-hoc test was used because it accounts for unequal sample sizes. The significance level was set to $p < 0.05$.

V. Results

A. Experiment 1: Leave-One-Subject-Out Models

The performance of ANN models validated using LOSO cross-validation within the datasets with raw data and tailored data (acceleration limited to ±5 g) are shown in Table III. In LOSO cross-validation, with both raw and tailored data, the overall classification accuracy of the hip- and wrist-based models were high, achieving above 80% (range: 81.8–95.4%), except for the dataset PAMAP (W), which showed a slightly lower classification accuracy (raw: 71.9%, tailored: 79.6%). The hip-based models exhibited an almost perfect agreement (range: K = 0.81–0.94) except in the dataset UOULU (H) when trained using raw data that exhibited substantial agreement (K = 0.78). The wrist-based models showed substantial agreement (K = 0.63–0.72) except in the dataset OSU (W) where the

TABLE III

CONFUSION MATRICES SHOWING THE PERFORMANCE OF ANN MODELS IN ACTIVITY INTENSITY CLASSIFICATION WITH RAW DATA AND TAILORED DATA (ACCELERATION LIMITED TO ±5 g) VALIDATED USING LEAVE-ONE-SUBJECT-OUT CROSS-VALIDATION WITHIN DATASETS

Validation: OULU (H)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	81.1	17.9	1.1	83.4 (78.1-88.8) Kappa, K: 0.78	SB	83.2	15.3	1.6	85.6 (80.2-91.0) Kappa, K: 0.81
LPA	9.1	85.5	5.5		LPA	9.7	86.7	3.6	
MVPA	1.2	17.8	81.1		MVPA	0.4	14.7	84.9	

Validation: OSU (H)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	98.0	0.6	1.4	94.4 (93.0-95.8) Kappa, K: 0.92	SB	98.6	0.3	1.1	95.4 (94.1-96.8) Kappa, K: 0.94
LPA	6.2	82.3	11.5		LPA	3.2	85.3	11.5	
MVPA	1.6	2.6	95.8		MVPA	0.7	2.7	96.6	

Validation: OSU (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	96.8	2.3	0.9	87.8 (83.8-91.9) Kappa, K: 0.85	SB	96.4	0.0	3.6	87.1 (81.5-92.6) Kappa, K: 0.84
LPA	4.8	58.1	37.1		LPA	8.1	57.3	34.7	
MVPA	2.0	4.4	93.6		MVPA	1.2	6.1	92.7	

Validation: PAMAP (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	66.4	30.0	3.6	71.9 (60.0-83.9) Kappa, K: 0.63	SB	68.2	12.7	19.1	79.6 (65.5-93.8) Kappa, K: 0.72
LPA	17.8	51.4	30.8		LPA	9.3	53.3	37.4	
MVPA	4.9	11.4	83.8		MVPA	2.2	4.9	93.0	

Validation: DSA (W)*

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	71.7	11.7	16.7	81.8 (73.6-89.9) Kappa, K: 0.66	SB	71.7	11.7	16.7	81.8 (73.6-89.9) Kappa, K: 0.66
LPA	2.5	48.1	49.4		LPA	2.5	48.1	49.4	
MVPA	1.1	3.4	95.6		MVPA	1.1	3.4	95.6	

The values across the intensity categories and overall accuracy (95% confidence interval) are presented in percentage (%). SB: Sedentary behavior, LPA: Light physical activity, MVPA: Moderate-to-vigorous physical activity. ∗The raw data and tailored data were similar.

agreement was almost perfect (raw data: K = 0.85, tailored data: K = 0.84). With respect to the three intensity categories, in all the datasets except UOULU (H), the LPA had the minimum classification accuracy, but the observations in SB and MVPA were mixed. Across all the five datasets, the differences in classification accuracy of the three intensity categories with raw and tailored data were marginal (range: 0.2–9.2 percentage points), resulting in marginal differences in overall classification accuracy (0.7–7.7 percentage points) and Kappa values (K = 0.01–0.09).

Table IV shows the performance of the models validated by the LOSO technique within a population when cross-tested on other populations with raw and tailored data. When the models were cross-tested with another dataset, the overall classification accuracies across the different datasets ranged from 41.2% to 59.9% and the Kappa values from K = 0.21 to 0.48. All models demonstrated lower performance than those obtained by within dataset cross-validation (Table III). The reduction in classification accuracy ranged from 20.1 to 44.3 percentage points. The Kappa values were also lower ranging from K = 0.24 to 0.63, resulting in fair or moderate agreement (K = 0.21–0.48). Raw and tailored data, even though in some cases favoring the classification accuracy of one or two of the intensity categories interchangeably, minimally affected the overall accuracies (0.7–13.1 percentage points) and Kappa values (K = 0.01–0.11).

TABLE IV

CONFUSION MATRICES SHOWING THE PERFORMANCE OF LEAVE-ONE-SUBJECT-OUT VALIDATED ANN MODELS IN ACTIVITY INTENSITY CLASSIFICATION IN AN INDEPENDENT POPULATION GROUP WITH RAW DATA AND TAILORED DATA (ACCELERATION LIMITED TO ±5 g)

Model: UOULU (H), Validation: OSU (H)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	49.6	46.3	4.1	46.5 (43.9-49.1) Kappa, K: 0.38	SB	66.2	32.9	0.9	52.2 (48.8-55.7) Kappa, K: 0.47
LPA	37.4	50.6	12.0		LPA	53.1	46.6	0.2	
MVPA	1.4	54.6	44.1		MVPA	2.8	49.8	47.5	

Model: OSU (H), Validation: UOULU (H)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	64.7	1.6	33.7	59.9 (54.1-65.7) Kappa, K: 0.48	SB	74.2	3.2	22.6	59.1 (55.1-63.1) Kappa, K: 0.47
LPA	31.5	30.0	38.5		LPA	27.9	37.6	34.5	
MVPA	1.2	5.0	93.8		MVPA	1.5	23.9	74.5	

Model: OSU (W), Validation: PAMAP (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	77.2	12.7	10.1	51.8 (43.2-60.5) Kappa, K: 0.39	SB	87.3	0.9	11.8	54.0 (47.4-60.6) Kappa, K: 0.44
LPA	38.4	23.9	37.7		LPA	43.0	13.1	43.9	
MVPA	3.8	32.4	63.8		MVPA	8.6	35.1	56.2	

Model: OSU (W), Validation: DSA (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	92.5	0.0	7.5	44.1 (39.5-48.7) Kappa, K: 0.25	SB	94.2	1.7	4.2	42.8 (37.2-48.4) Kappa, K: 0.21
LPA	58.1	10.0	31.9		LPA	52.5	9.4	38.1	
MVPA	28.2	28.4	43.4		MVPA	38.9	20.0	41.1	

Model: PAMAP (W), Validation: OSU (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	79.5	18.2	2.3	43.9 (39.0-48.8) Kappa, K: 0.30	SB	88.2	9.1	2.7	57.0 (48.9-65.1) Kappa, K: 0.41
LPA	35.5	49.2	15.3		LPA	58.1	15.3	26.6	
MVPA	24.4	57.3	18.3		MVPA	35.5	14.0	50.6	

Model: PAMAP (W), Validation: DSA (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	15.0	82.5	2.5	45.8 (41.2-50.4) Kappa, K: 0.21	SB	97.5	2.5	0.0	41.2 (34.8-47.6) Kappa, K: 0.22
LPA	18.1	48.8	33.1		LPA	86.9	3.8	09.4	
MVPA	8.2	39.2	52.6		MVPA	52.6	7.8	39.6	

Model: DSA (W), Validation: OSU (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	92.5	0.0	7.5	44.1 (39.5-48.7) Kappa, K: 0.25	SB	94.2	1.7	4.2	42.8 (37.2-48.4) Kappa, K: 0.21
LPA	58.1	10.0	31.9		LPA	52.5	9.4	38.1	
MVPA	28.2	28.4	43.4		MVPA	38.9	20.0	41.1	

Model: DSA (W), Validation: PAMAP (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy:	Class	SB	LPA	MVPA	Accuracy:
SB	28.2	25.5	46.4	49.1 (37.6-60.5) Kappa, K: 0.24	SB	28.2	25.5	46.4	49.8 (38.5-61.1) Kappa, K: 0.25
LPA	25.2	13.1	61.7		LPA	25.2	13.1	61.7	
MVPA	1.1	12.4	86.5		MVPA	1.1	10.8	88.1	

The values across the intensity categories and overall accuracy (95% confidence interval) are presented in percentage (%). SB: Sedentary behavior, LPA: Light physical activity, MVPA: Moderate-to-vigorous physical activity.

B. Experiment 2: Merging Various Sources of Data, Validation With Independent Dataset

The performance of the ANN models trained on merged training sets in the classification of activity intensities in another population that was not part of the training data are shown in Table V for both raw and tailored data. All possible dataset combinations were analyzed, having one population left out at a time, but only the most optimal results are reported here. The overall classification accuracy of the models trained on merged datasets ranged from 65.9 to 83.7% with substantial agreement (K = 0.61–0.79). In some cases, raw and tailored data favored the classification accuracy of one or two intensity categories interchangeably. However, the differences in overall accuracies (0.07–10.6 percentage points) and Kappa values (K = 0.01–0.11) remained marginal.

In all cases, the models trained on the merged datasets yielded a better generalization performance (Table V) compared to those obtained by placement-specific models (Table IV). This was primarily attributable to the better classification of all three intensity categories. The overall classification accuracy and agreements across the datasets were lower (2.5–19.3 percentage points, K = 0.01–0.21) but approached those obtained by within-dataset cross-validation (Table III) and for the dataset PAMAP (W) slightly higher to those obtained by within-dataset cross-validation (raw: 7.5 percentage points, K = 0.10; tailored: 4.1 percentage points, K = 0.06).

TABLE V

Confusion Matrices Showing the Performance of ANN Models Trained on Merged Datasets With Raw Data and Tailored Data (Acceleration Limited to ±5 g) in Activity Intensity Classification in an Independent Population

Model: OSU (H) + OSU (W) + PAMAP (W), Validation: UOULU (H)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 80.9 (77.8-84.0) Kappa, K: 0.77	Class	SB	LPA	MVPA	Accuracy: 70.3 (64.9-75.7) Kappa, K: 0.66
SB	92.6	5.3	2.1		SB	91.6	4.2	4.2	
LPA	25.5	66.1	8.5		LPA	35.5	43.3	21.2	
MVPA	2.3	7.7	90.0		MVPA	3.1	9.3	87.6	

Model: UOULU (H) + PAMAP (W) + DSA (W), Validation: OSU (H)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 82.9 (81.4-84.4) Kappa, K: 0.79	Class	SB	LPA	MVPA	Accuracy: 82.2 (79.7-84.6) Kappa, K: 0.79
SB	82.1	17.9	0.0		SB	87.5	12.5	0.0	
LPA	14.7	81.3	4.0		LPA	25.7	70.8	3.5	
MVPA	1.7	14.7	83.6		MVPA	1.3	16.0	82.7	

Model: UOULU (H) + PAMAP (W) + DSA (W), Validation: OSU (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 68.5 (65.1-71.9) Kappa, K: 0.64	Class	SB	LPA	MVPA	Accuracy: 71.0 (67.7-74.4) Kappa, K: 0.67
SB	62.3	37.7	0.0		SB	73.2	23.2	3.6	
LPA	0.0	33.9	66.1		LPA	0.0	50.8	49.2	
MVPA	1.2	14.5	84.3		MVPA	0.0	23.5	76.5	

Model: OSU (H) + OSU (W) + UOULU (H), Validation: PAMAP (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 79.4 (72.7-86.2) Kappa, K: 0.73	Class	SB	LPA	MVPA	Accuracy: 83.7 (76.9-90.4) Kappa, K: 0.78
SB	84.8	15.2	0.0		SB	90.9	4.5	4.5	
LPA	23.2	67.4	9.4		LPA	19.6	68.2	12.1	
MVPA	1.1	10.3	88.6		MVPA	1.1	10.3	88.6	

Model: OSU (H) + OSU (W) + UOULU (H), Validation: DSA (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 67.1 (63.9-70.4) Kappa, K: 0.69	Class	SB	LPA	MVPA	Accuracy: 71.1 (66.3-75.9) Kappa, K: 0.68
SB	76.7	23.3	0.0		SB	74.2	25.8	0.0	
LPA	3.1	63.8	33.1		LPA	6.9	63.1	30.0	
MVPA	0.2	33.9	65.9		MVPA	3.6	23.4	73.1	

The values across the intensity categories and overall accuracy (95% confidence interval) are presented in percentage (%). SB: Sedentary behavior, LPA: Light physical activity, MVPA: Moderate-to-vigorous physical activity.

C. Experiment 3: Merging Various Sources of Data, Validation Across all Datasets

The performance of the ANN models trained with merged data of 50% of the subjects in each dataset and validated with the remaining 50% of subjects is shown in Table VI. With both raw and tailored data, across all the five datasets, the model performed with over 80% classification accuracy (80.4–90.7%) and substantial or almost perfect agreement (K = 0.68–0.89). Across the five datasets, there were also small differences in the classification accuracy of the three intensity categories with raw and tailored data (0.0–19.2 percentage points), resulting in minor differences in overall classification accuracy (0.3–4.1 percentage points) and Kappa statistics (K = 0.01–0.04).

The sensitivity and specificity of the model developed with original raw data in classifying the three intensity categories across the five datasets are shown in Fig. 4. Since only minimal differences between raw and tailored were observed, the results of the tailored data are not shown here. Across all the datasets, the model classified SB and MVPA with both high (>80%) and comparable sensitivity except the SB in the UOULU (H) and OSU (W), which were classified by a sensitivity of slightly lower than 80% and significantly lower than OSU (H), and MVPA in DSA (W), which was over 80% but significantly lower than the datasets UOULU (H) and OSU (W). Similarly, the model also had high and comparable specificity in classifying the SB and MVPA across all the datasets except in OSU (W) and DSA (W) where the MVPA was significantly lower than OSU (H). For LPA, however, the model preformed with relatively lower sensitivity yet comparable across all the five datasets (∼60–80%). The specificity of classifying LPA was also high and comparable in the five datasets.

TABLE VI

Confusion Matrices Showing the Performance of ANN Models in Activity Intensity Classification With Raw Data and Tailored Data (Acceleration Limited to ±5 g), Trained on a Merged Dataset Consisting of 50% UOULU (h), 50% OSU (h), 50% OSU (w), 50% PAMAP (w) and 50% DSA (w) Subjects, and Validated With the Hold-Out Subjects

Validation: 50% UOULU (H)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 83.5 (75.7-91.3) Kappa, K: 0.78	Class	SB	LPA	MVPA	Accuracy: 80.4 (72.8-88.0) Kappa, K: 0.74
SB	74.4	24.4	1.1		SB	71.1	26.7	2.2	
LPA	7.5	80.1	12.4		LPA	11.2	75.2	13.7	
MVPA	2.5	4.2	93.3		MVPA	1.7	5.0	93.3	

Validation: 50% OSU (H)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 89.9 (87.9-91.8) Kappa, K: 0.88	Class	SB	LPA	MVPA	Accuracy: 90.7 (89.2-92.2) Kappa, K: 0.89
SB	97.2	2.8	0.0		SB	97.2	2.8	0.0	
LPA	8.0	76.0	16.0		LPA	6.5	79.0	14.5	
MVPA	1.4	8.5	90.1		MVPA	0.9	8.4	90.8	

Validation: 50% OSU (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 84.1 (80.0-88.2) Kappa, K: 0.82	Class	SB	LPA	MVPA	Accuracy: 83.0 (78.6-87.3) Kappa, K: 0.81
SB	78.6	19.6	1.8		SB	75.0	24.1	0.9	
LPA	1.6	60.9	37.5		LPA	3.1	68.8	28.1	
MVPA	0.0	4.0	96.0		MVPA	0.0	6.8	93.2	

Validation: 50% PAMAP (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 85.6 (74.4-96.7) Kappa, K: 0.77	Class	SB	LPA	MVPA	Accuracy: 81.5 (66.6-96.6) Kappa, K: 0.74
SB	83.3	8.3	8.4		SB	64.1	33.3	2.6	
LPA	10.7	73.2	16.1		LPA	14.6	68.3	17.1	
MVPA	1.0	5.2	93.8		MVPA	1.0	4.2	94.8	

Validation: 50% DSA (W)

Raw					Tailored				
Class	SB	LPA	MVPA	Accuracy: 80.7 (76.5-84.8) Kappa, K: 0.68	Class	SB	LPA	MVPA	Accuracy: 80.4 (74.0-86.9) Kappa, K: 0.70
SB	86.7	11.7	1.7		SB	85.0	11.7	3.3	
LPA	1.2	71.2	27.5		LPA	6.2	66.2	27.5	
MVPA	8.0	9.7	82.3		MVPA	3.8	12.2	84.0	

The values across the intensity categories and overall accuracy (95% confidence interval) are presented in percentage (%). SB: Sedentary behavior, LPA: Light physical activity, MVPA: Moderate-to-vigorous physical activity.

VI. Discussion

The purpose of this study was to evaluate and enhance the generalization performance of ML-based PA intensity prediction models to other populations monitored by different accelerometers from the one used in model development. The results demonstrated that merging data from hip and wrist accelerometers collected from various population groups monitored with different devices is a viable approach to enhance the generalization performance of ANN models in PA intensity classification across different population groups monitored by a single hip- or wrist-worn accelerometer. Instead, it seems that the high performance of LOSO-validated models is not generalizable to other population groups with accelerometers and characteristics different from the ones used to develop the models.

A. Leave-One-Subject-Out-Based Models and Their Generalization Performance

In LOSO cross-validation within each dataset, the hip- and wrist-based models with both raw data and tailored data with harmonized acceleration range showed high and comparable overall performance (Experiment 1), achieving over 80% accuracy (except for the PAMAP (W) dataset). The comparable overall performance of the models in activity intensity classification is consistent with previous studies that developed activity type classification models using hip and wrist accelerometer data [3], [15], [22]. With respect to the intensity categories, the LPA had the lowest classification accuracy across the datasets (except the dataset UOULU (H)), and the classification accuracy of the

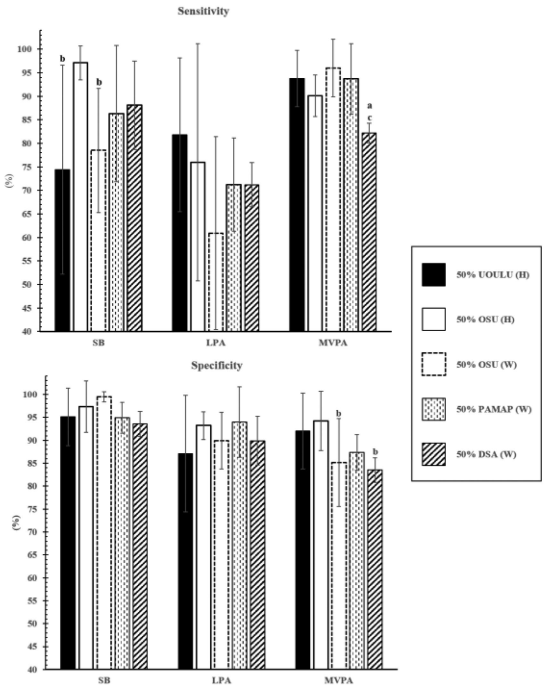

Fig. 4. Sensitivity and specificity of the ANN model with raw data in predicting sedentary behavior (SB), light physical activity (LPA), moderate-to-vigorous physical activity (MVPA) across the five datasets, trained on a merged dataset consisting of 50% UOULU (H), 50% OSU (H), 50% OSU (W), 50% PAMAP (W) and 50% DSA (W) subjects and validated with the hold-out subjects. a: significant difference compared to UOULU (H), b: significant difference compared to OSU (H), c: significant difference compared to OSU (W). Bars and the error bars represent mean values and their standard deviation.

other two categories were mixed. The relatively lower classification accuracy of LPA compared to SB and MVPA is consistent with past studies supporting the difficulty of distinguishing between LPA and lower- and higher-intensity categories [36]. According to a previous research showing the effects of data acquisition protocols on the performance of ML-based models [20], [37], the mixed results for the classification accuracy of SB and MVPA across the datasets with similar placement is perhaps due to differences in data acquisition protocols and performed activities in the datasets. This also may explain the higher accuracy of UOULU (H) in the detection of LPA, as the performed activities were mainly those with upper body movement. Similarly, the detailed observation of the datasets revealed that in PAMAP (W), the participants performed different sets of activities with a few overlapping activities, which might explain its low overall accuracy both across all the datasets and the wrist models. The performances of the PA intensity prediction models in the current study are similar to a recent one that validated ANN-based activity intensity predictions from various wear locations including hip and wrists [11].

When cross-validated to another population, the performance of all the models decreased, which was in line with previous studies [12], [20], [21], [38]. Compared to the achieved accuracy values in the previous studies (Table VII), our ANN models showed even more severe performance reduction using either raw or tailored data ranging from 20.1 to 44.3 percentage points. Recently, there has been a growing body of evidence showing that raw data acquired from different accelerometer brands or even models are incomparable, which can cause performance reduction when a model developed for one accelerometer is applied to another accelerometer [6], [7]. The differences in population characteristics and data acquisition protocols have also been identified as contributing factors to the predictive ability of models [12], [20], [21]. It appears that the more significant performance reduction of the ANN models might be due to the presence of multiple sources of heterogeneities across the datasets rather than merely one of them, including heterogeneity in data acquisition protocols (e.g., sensor attachment, monitored activities, etc.), population characteristics, and raw data (various activity monitors).

Consistent with previous findings on the generalization performance activity classification models [12], [20], [38], the results of this study highlight that the high accuracy of LOSO-validated models is not transferable to another population, and within-dataset cross-validation alone is not sufficient to understand how developed models will perform in a new population. Our results extend this finding by signifying that raw accelerometry and advanced modeling techniques do not necessarily warrant the generalizability of models on different populations, whose activities are monitored by different accelerometers. However, mainly due to the presence of several heterogeneities across the datasets and limited available meta-data from the open-access datasets, it is difficult to provide conclusive information about which factors played a more important role in the overall performance reduction of within-dataset-validated models. For the same reason, it remains elusive which factors to what extent caused a model trained on a certain dataset have a relatively better accuracy in classifying certain intensity categories in another dataset.

Minimal differences were seen in the performance of ANN models developed with raw or tailored data, when validated independently. It might be argued that rather than only limiting the acceleration range and sampling frequency, ascertaining the equivalency of raw data across different activity monitors using more sophisticated data conversion/filter strategies or extracting features that are comparable across monitors could have favored the generalization performance of the models. While this might be legitimate, it remains unclear how accelerometer outputs can be processed to provide comparable raw data, and controversies exist regarding comparable features across different accelerometers [1], [6], [39].

B. Merging Various Data Sources to Improve the Generalization Performance of the Models

The ANN activity intensity prediction models trained on a merged training set classified with acceptable performance the activity intensity in another population that was not part of the training phase and was monitored by a different accelerometer (Experiment 2). Their accuracies were 11.2 to 36.4 percentage points better than those obtained by the within-sample validated models, when applied to an independent population. The findings of accelerometry studies as well as in other fields of research can explain why the idea of merging data from multiple

TABLE VII
Achieved Accuracy Values and Characteristics of Populations in Studies Cross-Tested a Within-Sample Validated Model on an Independent Population

			Population used for model development with within-sample validation				Population used for independent validation of the within-sample validated model				
Study	Machine learning approach	Wear locati on	Number of Participants	Mean age± standard deviation/age range	Accelerometer sensor	**Accuracy**	Number of Participants	Mean age± standard deviation/age range	Accelerometer sensor	**Accuracy**	
[20]	Bayesian approach	Hip	59 (30 men, 29 females)	37.3±10.6/19-55	MotionLogs	**80.9%**	20 (NR)	NR/18–39	MotionLogs	**54.3%**	
[21]	Support vector machines (SVM)	Wrist	20 (12 boys, 8 girls)	13±1.3/11-15	Wocket	**91.0%**	33 (11 men, 22 females)	NR/18-75	Wocket	**71.8%**	
			33 (11 men, 22 females)	NR/18-75	Wocket	**87.0%**	20 (12 boys, 8 girls)	13±1.3/11-15	Wocket	**69.8%**	
[12]	Random forest	Wrist	39 (19 males, 20 females)	22.1±4.3/NR	GENEActiv	**92.8%**	24 (12 men, 12 females)	46.3±19.2/NR	GENEActiv	**77.3%**	
			24 (12 men, 12 females)	46.3±19.2/NR	GENEActiv	**80.2%**	39 (19 males, 20 females)	22.1±4.3/NR	GENEActiv	**78.5%**	
											Accuracy in the independent population using merged training sets (Table V) ⬇
This study*	Artificial nueral networks (ANN)	Hip	22 (11 males, 11 females)	27.5±11.2/17-58	Hookie AM20	**83.4%**	52 (28 boys, 24 girls)	13.7±3.1/7.2-18.9	ActiGraph GT3X+	**46.5%**	**80.9%**
			52 (28 boys, 24 girls)	13.7±3.1/7.2-18.9	ActiGraph GT3X+	**94.4%**	22 (11 males, 11 females)	27.5±11.2/17-58	Hookie AM20	**59.9%**	**82.2%**
		Wrist	16 (NR)	NR/NR	ActiGraph GT3X+	**87.8%**	9 (8 males, 1 females)	27.2±3.3/23-32	Colibri IMU	**51.8%**	**79.4%**
							8 (4 males, 4 females)	NR/20-30	Xsens IMU	**44.1%**	**67.1%**
			9 (8 males, 1 females)	27.2±3.3/23-32	Colibri IMU	**71.9%**	16 (NR)	NR/NR	ActiGraph GT3X+	**43.9%**	**68.5%**
							8 (4 males, 4 females)	NR/20-30	Xsens IMU	**45.8%**	**67.1%**
			8 (4 males, 4 females)	NR/20-30	Xsens IMU	**81.%**	16 (NR)	NR/NR	ActiGraph GT3X+	**44.1%**	**68.5%**
							9 (8 males, 1 females)	27.2±3.3/23-32	Colibri IMU	**49.1%**	**79.4%**

*The achieved accuracy values with downsampled raw acceleration data are presented. IMU: Inertial measurement unit, NR: Not reported.

population groups and different wear locations is plausible. The use of multiple accelerometers (e.g., hip and wrist) has shown improvements in the prediction accuracy of classification models [40]. For instance, classifying energy-consuming activities in which the wrist motion is limited (e.g., cycling) has been shown to be challenging with wrist accelerometers while the presence of motion from other wear locations (e.g., hip) seems to help classify such activities more accurately (or vice versa for other activities) [15]. While attaching multiple accelerometers might be cumbersome for the subjects and not feasible for real-world applications, combining data from various populations acquired from a single body-worn accelerometer may be more feasible for incorporating the information in data from various accelerometer placements. Previous studies have also demonstrated that training models on a wide range of activities along with the presence of interparticipant variability in age, body composition, and others can improve the generalization performance of ML-based models to a new population [12], [21]. Increased variability in raw acceleration data (data collected under different protocols or from different age groups) has been suggested to improve the generalization performance of ML-based models [20], [21], [37]. Indeed, according to demonstrations by other fields of research (e.g., speech recognition), the increased variability in population and raw data can be deemed as the addition of noise to the input data of an ANN during training, which results in significant improvements in generalization performance [41].

The results of this study suggest that integrating multiple datasets with data from only hip- or wrist-worn accelerometer into the training set might be a viable approach to augment the generalization performance of the models. This was supported by the performance enhancement of the ANN models close to the within-dataset validation performance or even better, when trained on merged datasets. The comparison between the achieved accuracy values in the present study and previous ones further indicates the robustness of the developed models using merged datasets (Table VII) even though the populations were monitored with different accelerometers and their characteristics remained different. The performance enhancements also support that integrating multiple datasets helps incorporate the information from accelerometers placed at the hip and wrist, together with increasing interparticipant and acceleration data variability. This finding is promising given that there has been a lack of methodologies for enhancing the robustness and generalization capability of ML models [4], [6], [12]. The marginal differences between raw and tailored data are also encouraging because they imply that enhancing the generalization performance of intensity prediction models can be done by combining original raw data even without data preprocessing (e.g., data filtering or conversion).

Finally, we also trained the model with a subset of subjects in all the datasets and validated it with the remaining subjects from all datasets (Experiment 3). This cross-study procedure resulted in a model with acceptable performance in classifying activity intensities across the datasets, achieving over 80% accuracy in all the five datasets. The model provided both comparable and high performance across the intensity categories, as evidenced by the high sensitivity and specificity of the model, with significant differences in only a few cases across five datasets. These results imply that a robust model that can predict activity intensities with both acceptable and comparable performance not only across different population groups with various accelerometers but also across hip- and wrist-worn accelerometers is achievable. This is new and noteworthy because it might have potential implications on enabling the comparability of accelerometry results.

Previous research has already shown several sources of heterogeneities that can independently or mutually affect the comparability of accelerometry data and subsequently the predictive performance of modeling approaches [6], [7], [20], [42]–[44]. To narrow down the gap between differences in accelerometry results, the existing literature has demanded a consensus on a single wear location [45]. However, even adopting this might not completely resolve the challenge since there are other sources of heterogeneities that are often inevitable (e.g., data acquisition protocol, sensor specifications, raw data from different accelerometers) and affect the comparability accelerometry data and modeling approaches (this was also shown in Experiment 1). Developing a robust generalizable model capable of operating acceptably in the presence of several heterogeneities might be a more feasible solution. It remains unclear how the in-lab prediction models will perform in free-living settings. Over longer periods, these nonsignificant differences in prediction performance of the model could lead to substantially different results. Independent cross-validation of the model is needed to ascertain that the model still provides comparable results under truly free-living settings and new populations.

C. Study Strengths and Limitations

The main strength of this study is the use of five different datasets. To the best our knowledge, this is the first cross-population study conducted on more than two populations, elucidating the generalization performance of PA prediction models. The heterogeneity of the datasets is also a strength; our study assumed that there would be heterogeneities in both accelerometry data and data collection protocols. This is not unrealistic, given that a variety of accelerometer-based activity monitors with different specification are now being designed that measure raw triaxial data, and researchers have started to obtain raw data with different protocols, parameters, and populations [5]. To facilitate generalizability and comparability between accelerometry results, our ANN models used the increased output comparability offered by raw accelerometry to provide more robust models rather than confining the measurement parameters, participants' age groups, and data collection protocols to certain and not-agreed-upon decisions.

This study has some limitations as well. Almost all the participants in the datasets PAMAP (W) and DSA (W) were right-handed, and for the dataset OSU (W), handedness was not specified. Due to the limited available meta-data regarding the detailed characteristics of participants in all the three sets of open-access data, it was not possible to eliminate the left-handed participants from the study. Using direct observation and the Compendium of Physical Activity as criterion measures to define activity intensities is also a limitation. This decision was made due to the lack of measured energy expenditure in the datasets and resulted in assigning all the activities of the same type to the same intensity category without considering the variability between the subjects in energy expenditure and performing a certain activity. Using similar MET thresholds for defining activity intensity categories in both adults and youths might also be a limitation. Previous studies on adults have consistently used the MET thresholds used in the present study to define intensity thresholds. However, there has been debate among calibration studies regarding the selection of MET intensity thresholds for youths [46]. We used the same value of ≥ 3 MET to define MVPA for both youths and adults to study the generalization performance of the models across the five datasets, which is an adapted intensity threshold for the both age groups in previous studies [13], [47].

VII. Conclusion

In conclusion, our cross-population study found that integrating heterogeneous datasets containing hip or wrist data in training sets is a viable approach to enhance the generalization performance of the ANN models as well as provide a model that predicts activity intensities across different populations with a single different hip- or wrist-worn accelerometers. Independent validation of within-sample validated models indicated that the ANN models developed with raw data and a within-population cross-validation technique (i.e., LOSO) are not generalizable to other populations monitored with different accelerometers. It seems that the performance deterioration might be mainly due to the presence of multiple sources of heterogeneities across datasets. Some of these heterogeneities, such as sensor specifications and population characteristics, are often inevitable and might not be resolved even through consensus on sensor placement, for example. The experiments further revealed that while the heterogeneities in datasets can adversely affect the generalization performance of intensity prediction models, if addressed properly in analytical approaches, they can be beneficial for improving the robustness of activity intensity prediction models. Our proposed method integrates various data sources in training sets to address heterogeneities in modeling level and train more robust models. However, it is still needed to confirm in future studies that the proposed method is capable of providing models that are robust enough to predict PA intensities in datasets acquired under fully free-living conditions with acceptable accuracies. The proposed method also seems a transparent, replicable, and feasible method, and can also be extendable assuming that the research lab will continuously share their data with proper meta-data. Sharing data seems essential to understand the effects of different inevitable heterogeneities on the results of ML-based models and to develop more robust models.

References

[1] D. R. Bassett Jr, A. V Rowlands, and S. G. Trost, "Calibration and validation of wearable monitors," *Med. Sci. Sports Exercise*, vol. 44, no. 1, pp. S32–S38, 2012.

[2] K. L. Cain, J. F. Sallis, T. L. Conway, D. Van Dyck, and L. Calhoon, "Using accelerometers in youth physical activity studies: A review of methods," *J. Phys. Activity Health*, vol. 10, no. 3, pp. 437–450, 2013.

[3] K. Ellis, J. Kerr, S. Godbole, J. Staudenmayer, and G. Lanckriet, “Hip and wrist accelerometer algorithms for free-living behavior classification,” *Med. Sci. Sports Exercise*, vol. 48, no. 5, pp. 933–940, 2016.

[4] M. de Almeida Mendes, I. C. M. da Silva, V. V Ramires, F. F. Reichert, R. C. Martins, and E. Tomasi, “Calibration of raw accelerometer data to measure physical activity: A systematic review,” *Gait Posture*, vol. 61, pp. 98–110, 2018.

[5] V. Farrahi, M. Niemelä, M. Kangas, R. Korpelainen, and T. Jämsä, “Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches,” *Gait Posture*, vol. 68, pp. 285–299, 2019.

[6] A. H. K. Montoye *et al.*, “Raw and count data comparability of hip-worn ActiGraph GT3X+ and link accelerometers,” *Med. Sci. Sports Exercise*, vol. 50, no. 5, pp. 1103–1112, 2018.

[7] D. John, J. Sasaki, J. Staudenmayer, M. Mavilia, and P. S. Freedson, “Comparison of raw acceleration from the GENEA and ActiGraph GT3X+ activity monitors,” *Sensors*, vol. 13, no. 11, pp. 14754–14763, 2013.

[8] C. E. Matthews, M. Hagströmer, D. M. Pober, and H. R. Bowles, “Best practices for using physical activity monitors in population-based research,” *Med. Sci. Sports Exercise*, vol. 44, no. 1, pp. S68–S76, 2012.

[9] K. Wijndaele *et al.*, “Utilization and harmonization of adult accelerometry data: Review and expert consensus,” *Med. Sci. Sports Exercise*, vol. 47, no. 10, pp. 2129–2139, 2015.

[10] P. Tjurin, M. Niemelä, M. Huusko, R. Ahola, M. Kangas, and T. Jämsä, “Classification of physical activities and sedentary behavior using raw data of 3D hip acceleration,” in *Proc. Joint Conf. Eur. Med. Biol. Eng. Conf. Nordic-Baltic Conf. Biomed. Eng. Med. Phys.*, 2017, pp. 872–875.

[11] A. H. K. Montoye, J. M. Pivarnik, L. M. Mudd, S. Biswas, and K. A. Pfeiffer, “Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior,” *AIMS Public Health*, vol. 3, no. 2, pp. 298–312, 2016.

[12] A. H. K. Montoye, B. S. Westgate, M. R. Fonley, and K. A. Pfeiffer, “Cross-validation and out-of-sample testing of physical activity intensity predictions using a wrist-worn accelerometer,” *J. Appl. Physiol.*, vol. 124, no. 5, pp. 1284–1293, 2018.

[13] J. Staudenmayer, S. He, A. Hickey, J. Sasaki, and P. Freedson, “Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements,” *J. Appl. Physiol.*, vol. 119, no. 4, pp. 396–403, 2015.

[14] A. H. K. Montoye, J. M. Pivarnik, L. M. Mudd, S. Biswas, and K. A. Pfeiffer, “Comparison of activity type classification accuracy from accelerometers worn on the hip, wrists, and thigh in young, apparently healthy adults,” *Meas. Phys. Educ. Exercise Sci.*, vol. 20, no. 3, pp. 173–183, 2016.

[15] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D. Wing, and S. Marshall, “A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers,” *Physiol. Meas.*, vol. 35, no. 11, pp. 2191–2203, 2014.

[16] D. M. Pober, J. Staudenmayer, C. Raphael, and P. S. Freedson, “Development of novel techniques to classify physical activity mode using accelerometers,” *Med. Sci. Sports Exercise*, vol. 38, no. 9, pp. 1626–1634, 2006.

[17] S. Zhang, A. V Rowlands, P. Murray, and T. L. Hurst, “Physical activity classification using the GENEA wrist-worn accelerometer,” *Med. Sci. Sport. Exercise*, vol. 44, no. 4, pp. 742–748, 2012.

[18] R. J. Kate, A. M. Swartz, W. A. Welch, and S. J. Strath, “Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data,” *Physiol. Meas.*, vol. 37, no. 3, pp. 360–379, 2016.

[19] S. G. Trost, W.-K. Wong, K. A. Pfeiffer, and Y. Zheng, “Artificial neural networks to predict activity type and energy expenditure in youth,” *Med. Sci. Sports Exercise*, vol. 44, no. 9, pp. 1801–1809, 2012.

[20] T. Bastian *et al.*, “Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: Laboratory-based calibrations are not enough,” *J. Appl. Physiol.*, vol. 118, no. 6, pp. 716–722, 2015.

[21] A. Mannini, M. Rosenberger, W. L. Haskell, A. M. Sabatini, and S. S. Intille, “Activity recognition in youth using single accelerometer placed at wrist or ankle,” *Med. Sci. Sports Exercise*, vol. 49, no. 4, pp. 801–812, 2017.

[22] S. G. Trost, Y. Zheng, and W.-K. Wong, “Machine learning for activity recognition: Hip versus wrist data,” *Physiol. Meas.*, vol. 35, no. 11, pp. 2183–2189, 2014.

[23] A. Reiss and D. Stricker, “Introducing a new benchmarked dataset for activity monitoring,” in *Proc. 16th Int. Symp. Wearable Comput.*, 2012, pp. 108–109.

[24] K. Altun, B. Barshan, and O. Tunçel, “Comparative study on classifying human activities with miniature inertial and magnetic sensors,” *Pattern Recognit.*, vol. 43, no. 10, pp. 3605–3620, 2010.

[25] “Publicly available datasets,” OSU Research Web site, 2011. [Online]. Available: http://web.engr.oregonstate.edu/∼wongwe/research.html. Accessed: Aug. 1, 2018.

[26] “PAMAP2 physical activity monitoring data set,” UCI Machine Learning Repository Web site, 2012. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/pamap2+physical+activity+monitoring. Accessed: Aug. 1, 2018.

[27] “Daily and sports activities data set,” UCI Machine Learning Repository Web site, 2013. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/daily+and+sports+activities. Accessed: Aug. 1, 2018.

[28] B. E. Ainsworth *et al.*, “2011 compendium of physical activities: A second update of codes and MET values,” *Med. Sci. Sport. Exercise*, vol. 43, no. 8, pp. 1575–1581, 2011.

[29] N. F. Butte *et al.*, “A youth compendium of physical activities: Activity codes and metabolic intensities,” *Med. Sci. Sport. Exercise*, vol. 50, no. 2, pp. 246–256, 2018.

[30] M. S. Tremblay *et al.*, “Sedentary behavior research network (SBRN) - terminology consensus project process and outcome,” *Int. J. Behavioral Nutrition Phys. Activity*, vol. 14, no. 75, pp. 1–17, 2017.

[31] S. Zhang, P. Murray, R. Zillmer, R. G. Eston, M. Catt, and A. V Rowlands, “Activity classification using the GENEA: optimum sampling frequency and number of axes,” *Med. Sci. Sport. Exercise*, vol. 44, no. 11, pp. 2228–2234, 2012.

[32] J. Staudenmayer, W. Zhu, and D. J. Catellier, “Statistical considerations in the analysis of accelerometry-based activity monitor data,” *Med. Sci. Sports Exercise*, vol. 44, no. 1, pp. S61–S67, 2012.

[33] J. E. Sasaki, A. Hickey, J. Staudenmayer, D. John, J. A. Kent, and P. S. Freedson, “Performance of activity classification algorithms in free-living older adults,” *Med. Sci. Sports Exercise*, vol. 48, no. 5, pp. 941–950, 2016.

[34] J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit,” *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968.

[35] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[36] L. D. Ellingson, I. J. Schwabacher, Y. Kim, G. J. Welk, and D. B. Cook, “Validity of an integrative method for processing physical activity data,” *Med. Sci. Sports Exercise*, vol. 48, no. 8, pp. 1629–1638, 2016.

[37] A. H. K. Montoye *et al.*, “Validation of accelerometer-based energy expenditure prediction models in structured and simulated free-living settings,” *Meas. Phys. Educ. Exercise Sci.*, vol. 21, no. 4, pp. 223–234, 2017.

[38] P. S. Freedson, K. Lyden, S. Kozey-Keadle, and J. Staudenmayer, “Evaluation of artificial neural network algorithms for predicting METs and activity type from accelerometer data: Validation on an independent sample,” *J. Appl. Physiol.*, vol. 111, no. 6, pp. 1804–1812, 2011.

[39] A. V Rowlands *et al.*, “Comparability of measured acceleration from accelerometry-based activity monitors,” *Med. Sci. Sports Exercise*, vol. 47, no. 1, pp. 201–210, 2015.

[40] S. G. Trost, D. Cliff, M. Ahmadi, N. Van Tuc, and M. Hagenbuchner, “Sensor-enabled activity class recognition in preschoolers: Hip versus wrist data,” *Med. Sci. Sports Exercise*, vol. 50, no. 3, pp. 634–641, 2017.

[41] S. Yin *et al.*, “Noisy training for deep neural networks in speech recognition,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2015, no. 2, pp. 1–14, 2015.

[42] M. E. Rosenberger, W. L. Haskell, F. Albinali, S. Mota, J. Nawyn, and S. Intille, “Estimating activity and sedentary behavior from an accelerometer on the hip or wrist,” *Med. Sci. Sports Exercise*, vol. 45, no. 5, pp. 964–975, 2013.

[43] A. H. Montoye, L. M. Mudd, S. Biswas, and K. A. Pfeiffer, “Energy expenditure prediction using raw accelerometer data in simulated free living,” *Med. Sci. Sport. Exercise*, vol. 47, no. 8, pp. 1735–1746, 2015.

[44] M. Hildebrand, V. T. H. VAN, B. H. Hansen, and U. L. F. Ekelund, “Age group comparability of raw accelerometer output from wrist-and hip-worn monitors,” *Med. Sci. Sports Exercise*, vol. 46, no. 9, pp. 1816–1824, 2014.

[45] S. J. Strath, K. A. Pfeiffer, and M. C. Whitt-Glover, “Accelerometer use with children, older adults, and adults with functional limitations,” *Med. Sci. Sports Exercise*, vol. 44, no. 1, pp. S77–S85, 2012.

[46] K. Ridley and T. S. Olds, “Assigning energy costs to activities in children: A review and synthesis,” *Med. Sci. Sport. Exercise*, vol. 40, no. 8, pp. 1439–1446, 2008.

[47] P. Freedson, D. Pober, and K. F. J. Janz, “Calibration of accelerometer output for children,” *Med. Sci. Sports Exercise*, vol. 37, no. 11, pp. S523–S530, 2005.