Gait & Posture 68 (2019) 285–299

Contents lists available at ScienceDirect

Gait & Posture

journal homepage: www.elsevier.com/locate/gaitpost

Review

Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches

Vahid Farrahi[a,*], Maisa Niemelä[a,b,c], Maarit Kangas[a,c], Raija Korpelainen[c,d,e], Timo Jämsä[a,b,c,f]

[a] Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu, Finland
[b] Infotech, University of Oulu, Oulu, Finland
[c] Medical Research Center, Oulu University Hospital and University of Oulu, Oulu, Finland
[d] Center for Life Course Health Research, University of Oulu, Oulu, Finland
[e] Oulu Deaconess Institute, Department of Sports and Exercise Medicine, Finland
[f] Diagnostic Radiology, Oulu University Hospital, Oulu, Finland

ARTICLE INFO

Keywords:
Objective measurement
Physical activity
Pattern recognition
Energy expenditure
Activity recognition

ABSTRACT

Background: Objective measures using accelerometer-based activity monitors have been extensively used in physical activity (PA) and sedentary behavior (SB) research. To measure PA and SB precisely, the field is shifting towards machine learning-based (ML) approaches for calibration and validation of accelerometer-based activity monitors. Nevertheless, various parameters regarding the use and development of ML-based models, including data type (raw acceleration data versus activity counts), sampling frequency, window size, input features, ML technique, accelerometer placement, and free-living settings, affect the predictive ability of ML-based models. The effects of these parameters on ML-based models have remained elusive, and will be systematically reviewed here. The open challenges were identified and recommendations are made for future studies and directions.
Method: We conducted a systematic search of PubMed and Scopus databases to identify studies published before July 2017 that used ML-based techniques for calibration and validation of accelerometer-based activity monitors. Additional articles were manually identified from references in the identified articles. Results: A total of 62 studies were eligible to be included in the review, comprising 48 studies that calibrated and validated ML-based models for predicting the type and intensity of activities, and 22 studies for predicting activity energy expenditure.
Conclusions: It appears that various ML-based techniques together with raw acceleration data sampled at 20–30 Hz provide the opportunity of predicting the type and intensity of activities, as well as activity energy expenditure with comparable overall predictive accuracies regardless of accelerometer placement. However, the high predictive accuracy of laboratory-calibrated models is not reproducible in free-living settings, due to transitive and unseen activities together with differences in acceleration signals.

1. Introduction

In research on physical activity (PA) and sedentary behavior (SB), precise assessment of PA and SB is essential in order to examine their relationship with various health outcomes, to assess the effectiveness of behavioral interventions, and to identify at-risk populations [1,2]. Over the past few years, objective measures, especially using wearable accelerometer-based activity monitors, have been extensively used for precise assessment of PA and SB [1–3]. However, robust data processing methodologies for converting accelerometer-provided data (signals) into different aspects of SB and PA, including time spent in various intensity categories, the type of activities performed, and estimate of energy expenditure (EE) have remained a non-trivial challenge [2–5].

Traditionally, accelerometers have provided proprietary manufacturer-provided measures known as “activity counts” (AC) and researchers have mostly used (linear) regression analysis to relate AC with activity energy expenditure (AEE) [2,5]. Receiver Operating Characteristics (ROC) curve analysis is another approach that has been used for establishing activity intensity cut-points [6]. Using these relatively simple methods (referred as traditional statistical methods in the literature), several regression-based equations and sets of thresholds (cut-points) have been established for estimating EE and activity intensity [7]. Regression-based equations and cut-points have been widely used in the existing literature for the assessment of PA and SB,

[*] Corresponding author at: Research Unit of Medical Imaging, Physics and Technology, P.O. 5000, FI-90014, University of Oulu, Finland.
E-mail address: Vahid.farrahi@oulu.fi (V. Farrahi).

https://doi.org/10.1016/j.gaitpost.2018.12.003
Received 30 July 2018; Received in revised form 8 November 2018; Accepted 3 December 2018

mainly due to their transparency and simplicity of implementation [7–9]. However, several studies have emphasized that these methods are not accurate in the presence of a wide range of activity types in free-living settings [2,3,5,10,11].

Therefore, there has been a demand for applying more sophisticated data modeling approaches [3,5,7,12,13]. There has been a shift from traditional statistical approaches towards ML-based modeling approaches, which are advanced statistical techniques with the ability to capture complicated relationships and nonlinearities in data, in order to calibrate and validate accelerometer-based activity data using either AC or raw acceleration data [2,6,7]. More recently, with the emergence of raw accelerometry, employment of more sophisticated data modeling approaches with raw acceleration data have been further emphasized [14]. It is widely accepted that raw accelerometry together with sophisticated modeling approaches, could enable developing precise data processing techniques [3,14]. Raw accelerometry has been also coupled with traditional statistical modeling approaches [6,15,16]. Hence, requests to discontinue the development of linear regression calibration equations and cut-points in order to avoid further methodological differences and discrepancies have emerged [3,5].

Despite the promising shift towards ML-based modeling approaches, several parameters and issues regarding the development and use of advanced ML-based modeling approaches have still remained elusive [2,3,5,13]. The effects of parameter choices, including data type (raw acceleration data versus AC), sampling frequency, window size, input features, axes of measurement, ML technique, and accelerometer placement, to predictive ability of ML-based models have still remained unclear [2]. Similarly, the predictive accuracy and generalization capability of ML-based models in relation to accelerometer placement in free-living settings is yet unknown [2,5,13].

Given the current existing knowledge gaps, we conducted a systematic review to identify studies that calibrated and validated accelerometer-based activity monitors using ML-based modeling approaches for predicting the type, intensity and/or EE of physical activity in order to examine (a) the reported effects of data type, sampling frequency, window size, input features, axes of measurement, and accelerometer placement on the predictive ability of ML-based models, and (b) the generalization capability of ML-based models to independent population in free-living settings. The main findings are further discussed in relation to the existing literature to further clarify how ML-based modeling approaches contribute to the existing analytical challenges of calibration and validation of accelerometer-based activity monitors. The recommendations for future studies and directions are made to move this field of research forward.

2. Methods

This systematic review is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [17]. Throughout this paper, the term “activity recognition” (AR) refers to all classification models calibrated and validated for classifying activities into types, classes, or intensities (i.e., using classification techniques not from metabolic equivalent [MET] predicted values) and the term “activity energy expenditure” (AEE) refers to all regression models calibrated and validated for estimating EE.

2.1. Eligibility criteria

We included original peer-reviewed journal articles including calibrated and validated ML-based models, based on data acquired from a single body-fixed accelerometer in order to predict the type, class, intensity and/or EE of activities. The monitored activities had to be health-related and daily activities such as walking, cycling, sedentary activities, etc. Studies validating a previously developed ML-based model were also included. In case of multiple accelerometers at various body locations the study was included if calibration and validation of ML-based models was based on data acquired from each attachment site. Studies with logistic regression (LR) were also included. Although LR is not always considered as a ML approach, it is a member of discriminative models in ML field.

Studies were excluded if they met at least one of the following exclusion criteria: (1) acceleration data gathered using smartphone-based or not body-fixed (e.g., the sensors were placed in subjects’ pockets) accelerometers; (2) studies using combined data acquired from multiple accelerometer placements (e.g., accelerometers placed at hip and wrist) or using combined data from combination of accelerometer data with other sensor data (e.g., gyroscope and accelerometers), in which no independent model using only acceleration data acquired merely from a single accelerometer placement was reported; (3) study participants had certain type of diseases (e.g., Parkinson’s, peripheral arterial disease, etc.) or disabilities; and (4) studies aimed at predicting activity type, class, intensity, and/or EE in real-time.

These exclusion criteria were selected because body-fixed accelerometers have been the most common method for objective measurement of SB and PA [2]. Most large-scale studies have used a single wear location to measure acceleration data [2]. Studies with smartphone-based or not body-fixed accelerometers were excluded to enable comparison between different wear locations. These studies make different assumptions such as changes in accelerometer’s position and orientation [18]. Studies with unhealthy subjects were excluded because the methods developed for healthy subjects might not be applicable to unhealthy subjects due to altered mobility and even different SB and PA indicators [5]. Finally, the studies aimed at measurement of SB and PA in real-time were excluded to enable fair comparison across the studies. These studies consider computational complexity of feature computation and prediction methods together with prediction accuracy for light-weight computation, which might result in sacrificing predictive accuracy (e.g., a limited number of features are extracted [19]).

2.2. Information sources

The PubMed and Scopus databases were initially searched on 01 July 2017. The two search strings are given in the Supplemental material, Search strings. The search results were limited for human subjects and English-language articles. The search results were not limited by date and included papers from all years in both databases. Additional articles were identified by searching the references in papers identified by the search strategy.

2.3. Study selection

Search strings were defined and validated by all authors (VF, MN, MK, RK, and TJ). One person (VF) performed the initial search and screened the titles and abstracts of all articles identified by the search strings to exclude articles that were not relevant according to the eligibility criteria. The initial literature search produced 3171 articles (Fig. 1). After removal of duplicate articles, 2482 articles remained for screening. On the basis of titles and abstracts screening, 2391 articles were excluded from the review. Additional 13 articles were manually identified from references in the papers. Two authors working independently (VF and MN) read the remaining 104 articles in full text and checked for eligibility using a predefined form including the eligibility criteria items. Disagreements concerning inclusion/exclusion of the papers were first resolved by discussion between VF and MN. The unresolved disagreements by discussion were resolved by a third independent reviewer (TJ). Of the 104 full text articles, 62 articles were finally considered eligible to be included in this review.

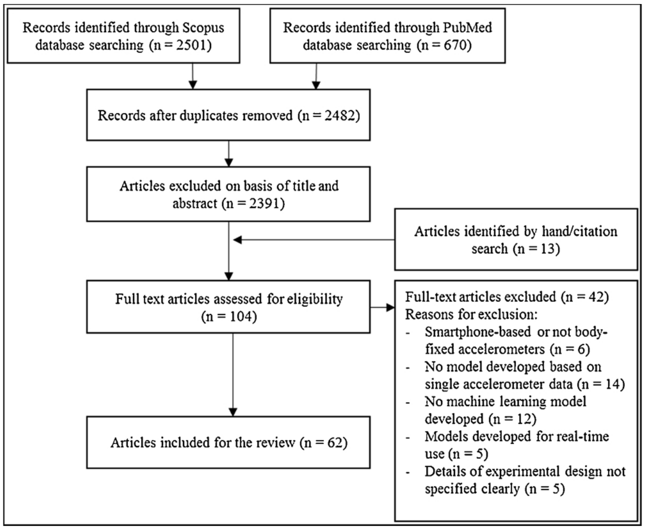
Records identified through Scopus database searching (n = 2501)
Records identified through PubMed database searching (n = 670)
Records after duplicates removed (n = 2482)
Articles excluded on basis of title and abstract (n = 2391)
Articles identified by hand/citation search (n = 13)
Full text articles assessed for eligibility (n = 104)
Full-text articles excluded (n = 42)
Reasons for exclusion:
- Smartphone-based or not body-fixed accelerometers (n = 6)
- No model developed based on single accelerometer data (n = 14)
- No machine learning model developed (n = 12)
- Models developed for real-time use (n = 5)
- Details of experimental design not specified clearly (n = 5)
Articles included for the review (n = 62)

Fig. 1. Flow diagram of selected studies.

2.4. *Risk of bias in studies*

The risk of bias at the study level was evaluated using the modified QualSyst tool [20]. The irrelevant criteria were excluded and some of them were adopted based on earlier recommendations, including representative activities and sample population, as well as a clear specification of extracted features, input features, and signal axes [2,13,21]. Finally, ten criteria were used for scoring the studies, which are presented in Table S1 in the Supplemental Material. The risk of bias was evaluated with a summary score (range: 0–1), with a higher score indicating better quality. The ten criteria were validated by all the authors and risk of bias was assessed by the first author (VF).

2.5. *Synthesis of the results*

Since classification (i.e., AR) and regression (i.e., AEE) are considered as two independent tasks in ML field [22], the studies calibrating and validating both AR and AEE models were counted and reported independently in both categories. To report the extracted data and their effects on results of ML-based models, if multiple parameters were tested (e.g., multiple window size), all the tested parameters were extracted from the included studies. The most optimal parameters as stated by the authors are reported as the parameters used for model development, whereas the effects of other tested parameters are discussed narratively in the text. Similarly, to report the predictive accuracies of ML-based models, overall accuracy (defined as the proportion of the total number of predictions that were correct) for AR models and root mean squared error (rMSE in MET values) for AEE models, achieved by the most optimal models were extracted and reported. For AR models, if overall accuracy was not presented, sensitivity (defined as the average of true positive rates for the activity categories) was extracted and reported. When the results were evaluated based on the extracted predictive accuracies, those studies which did not present overall accuracy or sensitivity for AR models and rMSE for AEE models were eliminated from the analysis.

Four common accelerometer placements were observed; including hip, wrist, ankle and thigh. The predictive accuracies were extracted in relation to these four placements. Additionally, the predictive accuracy of calibrated and validated models for any other wear locations were also extracted and reported under category ‘other placements’. If models were developed for more than one other wear location, the highest predictive accuracy achieved was extracted and reported under ‘other placements’ category. Similarly, if different anatomical sides were considered for accelerometer placement (i.e., left and right or dominant and non-dominant), the highest predictive accuracy achieved was extracted and reported. If more than one validation approach was used, the accuracies of leave-one-subject-out (LOSO) validation were extracted, which is known to provide a good estimate of model accuracy [3,23].

3. Results

3.1. *Quality of the included studies*

The mean quality score in the included studies was high (average: 0.88, range: 0.60–1.00). Main sources of bias were small sample size and monitoring a limited set of activity types. Quality scores of the studies in relation to the adopted criteria are presented in Table S1 in the Supplemental Material.

3.2. *The extracted data items and model type*

The extracted data items from the included studies along with population characteristics are presented in Table 1. Forty out of the 62 included studies calibrated and validated AR models, 14 calibrated and validated AEE models, and 8 calibrated and validated both AR and AEE models. Thus, the denominator is 48 when reference is made to AR studies/models and 22 when reference is made to AEE studies/models.

Fig. 2 and Fig. 3 show the extracted overall accuracies of the calibrated and validated AR and AEE models in relation to accelerometer placements and independent validation, respectively. In Fig. 2, sensitivity is presented as the measure of performance for two studies, since the overall accuracies were not presented [24,25]. Also, the results of three studies are not shown, since other measures of performance than overall accuracy were reported visually per activity, not as a single value [26–28]. In Fig. 3, median rMSEs or rMSE are estimated from the figures in the articles for three studies [29–31]. Additionally, the results of four studies are not shown because one of the studies examined the validity of an AEE model independently in relation to direct observation (not AEE) [32] and three studies presented other measures of performance than rMSE [33–35].

3.3. *Study characteristics*

3.3.1. *Sample population*

Eight studies were performed with more than one population. Sixteen populations within 15 studies comprised exclusively children and/or adolescents, 36 populations used within 34 studies comprised exclusively adults, and three populations comprised exclusively older adults. Additionally, fifteen populations within 14 studies comprised subjects with relatively wide age range (mixed age range) including children, adolescents, adults, and/or older adults. The age range was not reported in one study.

3.3.2. *Accelerometer placement*

The models were mainly developed for hip- (n = 43), wrist- (n = 31), ankle- (n = 15), and thigh-worn (n = 10) accelerometers, respectively. Other accelerometer placements (e.g., chest ear, knee, shin), altogether, were used in fourteen studies.

3.3.3. *Machine learning algorithm*

The artificial neural network (ANN) was the most commonly used ML algorithm in the included studies (n = 32), followed by support vector machines (SVM) (n = 18), random forest (RF) (n = 12), decision tree (DT) (n = 11), and LR (n = 7). Additionally, other ML algorithms including Naïve Bayes (NB), Bayesian network, K-nearest neighbors (KNN), and hidden Markov model (HMM), altogether, were used in seventeen studies.

Table 1
Population characteristics and the extracted data items for activity recognition and energy expenditure estimation studies, listed by first author's last name (publication year)).

Author (year)	Sample population size (male/female)	Mean age (years) (SD)/age range	Mean body mass index (SD)/range	Sensor brand and model	Sensor placement	Sampling frequency or activity counts	Window size	Feature selection; feature extraction	Signal axes used for feature extraction	Machine learning algorithm
Activity recognition										
Pober (2006) [36]	6 (4/2)	24.8 (4.2)/NR	NR/NR	ActiGraph 7164	Right hip	Activity counts	Non-overlapping 15 sec	NA; TD features	Uniaxial (y-axis)	HMM*, Quadratic discriminant analysis
Bonomi (2009) [37]	20 (13/7)	29.0 (6.0)/NR	23.6 (3.2)/NR	Tracmor	Waist (lower back)	20 Hz	Non-overlapping 0.4, 0.8, 1.6, 3.2, 6.4* and 12.8* sec	NA; TD and FD features	(x, y, z)	DT
Preece (2009) [38]	20 (10/10)	31.0 (7.0)/NR	24.0 (3.0)/NR	Pegasus	Hip, ankle, thigh	64 Hz	2 sec with 50% sec overlap	NA; TD*, FD*, and wavelet features	(x, y, z)	KNN
Khan (2010) [39]	6 (3/3)	27 (NR)/NR	NR/NR	Witilt	Chest	20 Hz	Non-overlapping 3.2 sec	NA; TD features	(x, y, z)	ANN
Atallah (2011) [34]	11 (9/2)	NR/NR	NR/NR	ADXL330, e-AR (ear-worn)	Hip, wrist, ankle, knee, arm, right ear, chest	50 Hz	Non-overlapping 5 sec	Relied, Simba, mRMR; TD and FD features	(x, y, z)	BN, KNN
De Vries (2011) [40]	49 (21/28)	38.0 (11.0)/22-62	23.8 (3.4)/NR	ActiGraph GT1M	Right hip and ankle	Activity counts	Non-overlapping 10 sec	Correlation-based; TD features	Uniaxial (y-axis)	ANN
De Vries (2011) [41]	58 (31/27)	11.0 (0.7)/9-12	19.0 (3.0)/NR	ActiGraph GT3X and GT1M	Hip, ankle	Activity counts	Non-overlapping 10 sec	NA; TD features	(x, y, z)	ANN
Gyllensten (2011) [42]	52 (29/23)	29.2 (6.7)/23-43	23.6 (3.2)/19.2-32.6	Tracmor	Waist (lower back)	20 Hz	Non-overlapping 6.4 sec	NA; TD and FD features	(x, y, z)	ANN, DT, SVM*, majority voting*
	20 (10/10)	30.0 (9.0)/22–51	23.0 (2.6)/18.8-28.0	Tracmor	Waist (lower back)	20 Hz				
Lee (2011) [43]	20 (NR)	NR/22-30	NR/NR	SerAccel	Chest	20 Hz	10 sec with 50% overlap	NA; TD features	(x, y, z, VM)	ANN
Ruch (2011) [44]	41 (17/24)	10.6 (1.6)/NR	NR/NR	Actigraph GT1M	Hip, wrist	Activity counts	1 sec epoch without feature extraction	NA; NA	Uniaxial (y-axis)	KNN
Schmid (2011) [45]	10 (6/4)	10 (6/4)	10 (6/4)	10 (6/4)	10 (6/4)	10 (6/4)	10 (6/4)	10 (6/4)	10 (6/4)	10 (6/4)
Oudre (2012) [24]	24 (15/9)	24.8 (12)/19-54	25.4 (4.4)/ 19.2-33.6	MotionPod	Shin	100 Hz	10.24 sec with 75% overlap	NA; FD features	z	Wasserstein distance
Zhang (2012) [46]	60 (23/37)	49.6 (6.4)/40-65	23.8 (3.5)/NR	GENEA	Right wrist	5, 10*, 20*, 40, and 80 Hz	Non-overlapping 12.8 sec	NA; TD, FD, and wavelet features	VM calculated from x*, y*, z*, (x, y), (x, z), (y, z), and (x, y, z) axes.	LR, DT, SVM, BN
Zhang (2012) [47]	60 (23/37)	49.4 (6.5)/40-65	24.6 (3.4)/NR	GENEA	Right hip, Left and right wrists	80 Hz	Non-overlapping 12.8 sec	SVM-based; TD and FD features	VM	ANN, DT*, LR, NB, SVM*
Cleland (2013) [48]	8 (8/0)	26.2 (2.8)/24-33	NR/NR	Shimmer	Left hip, wrist, thigh, and ankle, lower back, chest	51.2 Hz	10.24 sec with 50% overlap	NA; TD and FD features	(x, y, z)	DT, NB, ANN, SVM*
Hees (2013) [26]	28 (13/15)	Male: 31.8 (6.5), female: 29.8 (17.1)/21-53	Male: 22.7(1.2)/NR, female: 23.6(6.3)/NR	GENEA	Left and right, hips, wrists, and ankles, lower back, upper arm, upper leg	80 Hz	Non-overlapping 2 sec	PCA-based; TD and FD features	(x, y, z, VM)	LR
John (2013) [49]	10 (NR)	23.8 (5.4)/NR	22.7 (1.4)/NR	GENEA, ActiGraph GT3X+	Dominant wrist	80 Hz	Non-overlapping 20 sec	NA; TD and FD features	VM	RF
Mannini (2013) [50]	33 (11/22)	NR/18-75	NR/NR	Wockets	Dominant wrist, ankle	90 Hz		8 manually-provided feature	VM	SVM

(*continued on next page*)

Table 1 (*continued*)

Author (year)	Sample population size (male/ female)	Mean age (years) (SD)/age range	Mean body mass index (SD)/range	Sensor brand and model	Sensor placement	Sampling frequency or activity counts	Window size	Feature selection; feature extraction	Signal axes used for feature extraction	Machine learning algorithm
							Non-overlapping 2, 4, and 12.8* sec	sets; TD*, FD*, and wavelet features		
Zhao (2013) [51]	69 (NR)	NR/3-5	25% overweight or obese/NR	ActiGraph GT3X+	Right hip	Activity counts	Non-overlapping 60 sec	NA; TD features	(x, y, z)	LR, SVM*
He (2014) [28]	16 (8/8)	80.6 (4.8)/NR	26.1 (2.5)/NR	ActiGraph GT3X+	Right hip, left and right wrists	80 Hz	Non-overlapping 1 sec	NA; NA (distances between raw data)	(x, y, z)	Movelets (distance based)
Trost (2014) [52]	52 (28/24)	13.7 (3.1)/7.2-18.9	BMI percentile 50.2 (23.1)/ 4.6-89.9	ActiGraph GT3X+	Right hip, non-dominant wrist	30 Hz	Non-overlapping 6 sec	NA; TD-based features	NR	LR
Arif (2015) [53]	9 (8/1)	27.2 (3.3)/NR	25.11 (2.6)/NR	Colibri IMU	Dominant wrist and ankle, chest	100 Hz	5 sec with 20% overlap	Correlation-based; TD features	(x, y, z)	ANN, KNN, Rotation forest*
Bastian (2015) [54]	59 (30/29)	59 (30/29)	59 (30/29)	59 (30/29)	59 (30/29)	59 (30/29)	59 (30/29)	59 (30/29)	59 (30/29)	59 (30/29)
	20 (NR)	NR/18-39	NR/NR	MotionLogs	Hip	25 Hz				
Fida (2015) [55]	9 (4/5)	NR/22-34	NR/NR	Internally developed IMU	Waist	100 Hz	Non-overlapping 0.5, 1, 1.5*, 2, 2.5, and 3 sec	NA; TD features	(x, y, z)	DT, KNN, SVM, NB, ANN
Hagenbuchner (2015) [56]	11 (6/5)	4.8 (0.87)/3-6	15.9 (1.0)/NR	ActiGraph GT3X+	Hip	100 Hz converted to activity counts	Non-overlapping 10, 15, 20, 30 and 60* sec	NA; TD features	y	ANN feed-forward and self-organizing maps, deep learning ensemble network*
	100 (NR)	11.0 (2.7)/5-15	NR/NR	ActiGraph GT1M	Right hip	Activity counts				
Ellis (2016) [57]	40 (0/40)	55.2 (15.3)/NR	32.0 (3.7)	ActiGraph GT3X+	Right hip, non-dominant wrist	30 Hz	Non-overlapping 60 sec	NA; TD and FD features	VM	Two-layer RF and HMM
Kerr (2016) [58]	25 (18/7)	36.0 (12.0)/18-70	25% obese/NR	ActiGraph GT3X+	Right hip	30 Hz	Non-overlapping 60 sec	NA; TD and FD features	VM	Two-layer RF and HMM
	100 (0/100)	55.0 (16.0)/NR	100% obese/ NR	ActiGraph GT3X+	Right hip	30 Hz				
Margarito (2016) [25]	48 (24/24)	29.0 (9.0)/NR	27.0/NR	Philips DirectLife	Wrist	20 Hz	Non-overlapping 6 sec	NA; TD and FD features	VM	ANN*, DT, LR, NB, Template matching
Montoye (2016) [59]	40 (19/21)	22.0 (4.2)/18-44	24.3(3.5)/NR	GENEActiv, ActiGraph GT3X+	Right hip, right and left wrists, right thigh	20 Hz	Non-overlapping 30 sec	NA; TD features	(x, y, z)	ANN
Montoye (2016) [60]	39 (19/20)	22.1 (4.3)/18-44	24.4 (3.6)/NR	GENEActiv, ActiGraph GT3X+	Right hip, right and left wrists, right thigh	30 Hz	Non-overlapping 5 sec	4 manually-provided feature sets; TD features	(x, y, z)	ANN
Ren (2016) [61]	112 (58/54)	NR/8-12	NR/NR	ActiGraph GT3X	Right hip	Activity counts	Non-overlapping 60 sec	NA; TD and FD features	NR	DML-KNN (a distance based learning method)
Sasaki (2016) [62]	35 (14/21)	70.6 (5.0)/NR	26.8 (4.3)/NR	ActiGraph GT3X+	Dominant hip wrist, and ankle	80 Hz	Non-overlapping 5, 10, 20, and, 30* sec	NA; TD and FD features	(x, y, z)	RF, SVM
Arif (2017) [63]	9 (8/1)	9 (8/1)	9 (8/1)	9 (8/1)	9 (8/1)	9 (8/1)	9 (8/1)	9 (8/1)	9 (8/1)	9 (8/1)
Chowdhury (2017) [64]	9 (8/1)	27.2 (3.3)/NR	25.1 (2.6)/NR	Colibri IMU	Dominant wrist	100 Hz	10.24 sec with 50% overlap	Correlation based; TD and FD features	(x, y, z)	ANN, bagging, boosting, and binary DT, majority voting*, KNN, RF*, SVM
	8 (4/4)	29.9 (4.2)/NR	22.8 (1.9)/NR	Empatica E4	Non-dominant wrist	32 Hz				
	17 (9/8)	14.6 (2.4)/NR	BMI percentile: 66.8 (25.6)/NR	ActiGraph GT3X+	Non-dominant wrist	30 Hz				
Kühnhausen (2017) [65]	70 (NR)	9.77 (0.62)/8-11	NR/NR	ActiGraph GT3X+	Non-dominant hip	30 Hz*, activity counts	Non-overlapping 2.5 sec	NA; TD and FD features	(x, y, z)	SVM

(*continued on next page*)

Table 1 (*continued*)

Author (year)	Sample population size (male/female)	Mean age (years) (SD)/age range	Mean body mass index (SD)/range	Sensor brand and model	Sensor placement	Sampling frequency or activity counts	Window size	Feature selection; feature extraction	Signal axes used for feature extraction	Machine learning algorithm
Mannini (2017) [66]	20 (12/8)	13 (1.3)/NR	NR/NR	Wocket	Dominant wrist and ankle	90 Hz	Non-overlapping 3.2, 6.4, 8, 12.8* sec	Sequential forward search; TD and FD features	VM	SVM
	33 (11/22)	NR/18-75	NR/NR	Wocket	Dominant wrist and ankle	90 Hz				
Pavey (2017) [67]	21 (13/8)	27.6 (6.2)/NR	NR/NR	GENEActiv	Non-dominant wrist	30 Hz	Non-overlapping 5 sec	NA; TD and FD features	(x, y, z)	RF
Rosenberg (2017) [68]	39 (0/39)	69.4 (NR)/56-94	NR/19.7-45.6	ActiGraph GT3X+	Right hip	30 Hz	Non-overlapping 60 sec	NA; TD and FD features	VM	Two-layer RF and HMM
Trost (2017) [69]	11 (5/6)	4.8 (0.87)/3-6	15.9 (1.0)/NR	ActiGraph GT3X+	Right hip, non-dominant wrist	100 Hz	Non-overlapping 15 sec	NA; TD and FD features	VM	RF, SVM
Montoye (2018) [70]	39 (19/20)	22.1(4.3)/18-35	24.4 (3.6)/NR	GENEActiv	Left wrist	20 Hz	Non-overlapping 30 sec	6 manually-provided feature sets; TD* and FD features	(x, y, z)	ANN, DT, RF*, SVM, majority voting
	24 (12/12)	46.3 (19.2)/18-79	26.1 (3.6)/NR	GENEActiv	Left wrist	20 Hz				
Activity energy expenditure										
Rothney (2007) [35]	102 (46/56)	38.6 (13.1)/18-70	26.0 (5.3)/ 16.9–42.1	Custom-designed IDEEA	Right hip	32 Hz	Non-overlapping 60 sec	NA; TD and FD features	Biaxial	ANN
Atallah (2011) [27]	25 (18/7)	29.9 (4.5)/NR	NR/NR	e-AR	Ear	100 Hz	Non-overlapping 60 sec	KNN-based; TD and FD features	(x, y, z)	AdaBoost
Ruch (2013) [71]	43 (21/22)	9.8 (2.4)/NR	NR/NR	ActiGraph GT3X	Right hip	64 Hz	Non-overlapping 60 sec	NA; Activity counts as features	(x, y, z)	ANN
Lyden (2014) [72]	13 (5/8)	24.8 (5.2)/18-60	24.8 (5.2)/18-60	24.8 (5.2)/18-60	24.8 (5.2)/18-60	24.8 (5.2)/18-60	24.8 (5.2)/18-60	24.8 (5.2)/18-60	24.8 (5.2)/18-60	24.8 (5.2)/18-60
Altini (2015) [30]	15 (11/4)	29.8 (5.2)/NR	23.2 (3.0)/NR	ADXL330	Right hip, dominant ankle, wrist, and thigh, chest	64 Hz	Non-overlapping 4 sec	Correlation-based; TD and FD features	(x, y, z)	SVM
Kim (2015) [32]	11 (8/3)	30.6 (7.2)/NR	25.3 (4.5)/NR	ActiGraph GT3X	Right hip and thigh	Activity counts	NA (varying window length)	NA; TD features	y, (x, y, z)	Sojourn (Hand built decision tree and ANN)
Montoye (2015) [73]	39 (19/20)	22.1 (4.3)/18-44	24.4 (3.6)/NR	GENEActiv, ActiGraph GT3X+	Right hip and thigh, right and left wrists	20 Hz (GENEActiv), 40 HZ (ActiGraph)	Non-overlapping 60 sec	4 manually-provided feature sets; TD features	(x, y, z)	ANN
Ellingson (2016) [33]	49 (18/31)	23.9 (5.3)/18-40	23.4 (3.5)/NR	ActiGraph GT3X+	Right hip and thigh	100 Hz converted to activity counts	NA (varying window length)	NA; TD features	(x, y, z)	Sojourn 3x (Hand built decision tree and ANN)
Mackintosh (2016) [29]	27 (15/12)	10.8 (1.0)/NR	NR/NR	ActiGraph GT3X+	Right and left hips, wrists, ankles, and knees, chest	100 Hz converted to activity counts	Non-overlapping 15 sec	NA; TD features	(x, y, z)	ANN
Montoye (2016) [74]	25 (11/14)	21.8 (2.5)/18-30	23.0 (2.5)	ActiGraph GT3X+	Right hip	30 Hz	Non-overlapping 30 sec	NA; TD features	(x, y, z)	ANN
Montoye (2016) [75]	39 (19/20)	22.1 (4.3)/18-44	24.4 (3.6)/NR	GENEActiv	Left and right wrists	20 Hz	Non-overlapping 30 sec	3 manually-provided feature sets; TD features	(x, y, z, VM)	ANN
Montoye (2017) [76]	41 (20/21)	22.0 (4.2)/18-35	NR/NR	ActivPAL3	Right thigh	20 Hz	Non-overlapping 30 sec	NA; TD features	(x, y, z)	ANN
Montoye (2017) [31]	40 (19/21)	22.0 (4.2)/18-44	24.3 (3.5)/NR	GENEActiv, ActiGraph GT3X+	Right hip, left and right wrists, right thigh	20 Hz (GENEActiv), 40 HZ (ActiGraph)	Non-overlapping 30 sec	NA; TD features	(x, y, z)	ANN
Montoye (2017) [77]	24 (12/12)	45.8 (19.4)/18-80	26.1 (3.5)/NR	ActiGraph GT9X Link	Right hip and ankle, right and left wrists	60 Hz	Non-overlapping 30 sec	NA; TD features	(x, y, z)	ANN

(*continued on next page*)

Table 1 (*continued*)

Author (year)	Sample population size (male/female)	Mean age (years) (SD)/age range	Mean body mass index (SD)/range	Sensor brand and model	Sensor placement	Sampling frequency or activity counts	Window size	Feature selection; feature extraction	Signal axes used for feature extraction	Machine learning algorithm
Activity recognition and energy expenditure										
Staudenmayer (2009) [78]	48 (24/24)	35.0 (NR)/21–69	24.2(NR)/17.9-40.5	Actigraph 7164	Right hip	Activity counts	Non-overlapping 60 sec	NA; TD features	Uniaxial (y-axis)	ANN
Freedson (2011) [79]	277 (139/138)	38.0 (12.4)/NR	24.6 (4.0)/NR	ActiGraph GT1M	Non-dominant hip	Activity counts	Non-overlapping 60 sec	NA; TD features	Uniaxial (y-axis)	ANN
	65 (38/27)	40.1 (13.0)/NR	27.1 (5.6)/NR	ActiGraph GT1M	Non-dominant hip	Activity counts				
Trost (2012) [80]	100 (NR)	11.0 (2.7)/5-15	NR/NR	ActiGraph GT1M	Right hip	Activity counts	Non-overlapping 10, 15, 20, 30 and 60* sec	NA; TD features	Uniaxial (y-axis)	ANN
Ellis (2014) [81]	40 (19/21)	35.8 (12.1)/NR	24.8 (2.9)/NR	ActiGraph GT3X+	Left and right hip, non-dominant wrist	30 Hz	Non-overlapping 60 sec	NA; TD and FD features	(x, y, z, VM)	RF
Mu (2014) [82]	112 (57/55)	11.0 (1.7)/8-15	21.4 (5.5)/NR	ActiGraph GT3X	Right hip	Activity counts	Non-overlapping 60 sec	NA; TD features	(x, y, z)	ANN, Bipart (distance based)*, citation KNN, KNN, NB, SVM
Staudenmayer (2015) [83]	20 (10/10)	24.1 (2.9)/NR	23.9 (2.9)/NR	ActiGraph GT3X	Right hip, dominant wrist	80 Hz	Non-overlapping 15 sec	NA; TD features	VM	ANN, RF, SVM
Strath (2015) [84]	99 (48/51)	49.0 (17.4)/18-65 +	Body fat: 27.4 (10.0)/NR	ActiGraph GT3X	Non-dominant hip, wrist, and ankle	Activity counts	Non-overlapping 60 sec	NA; Ngram-based features	VM	SVM
Kate (2016) [85]	146 (67/79)	49.1 (17.3)/18-79	25.7 (4.4)/NR	ActiGraph GT3X	Non-dominant hip	Activity counts	Non-overlapping 60 sec	Several manually-provided feature sets; TD*, distance-, and Ngram-based features	VM	ANN, bagged DT, DT, KNN, LR, NB, RF, SVM

* Indicates optimal parameters mentioned by the authors. Abbreviations: NA: Not applied, NR: Not reported, BMI: Body mass index, IMU: Inertial measurement unit, ANN: Artificial neural network, BN: Bayesian network, DT: Decision tree, HMM: Hidden Markov models, KNN: K-nearest neighbors, LR: Logistic regression, NB: Naïve Bayes, PCA: Principal Component Analysis, RF: Random forest, SVM: Support vector machines, VM: Vector magnitude, TD: Time-domain, FD: Frequency-domain.

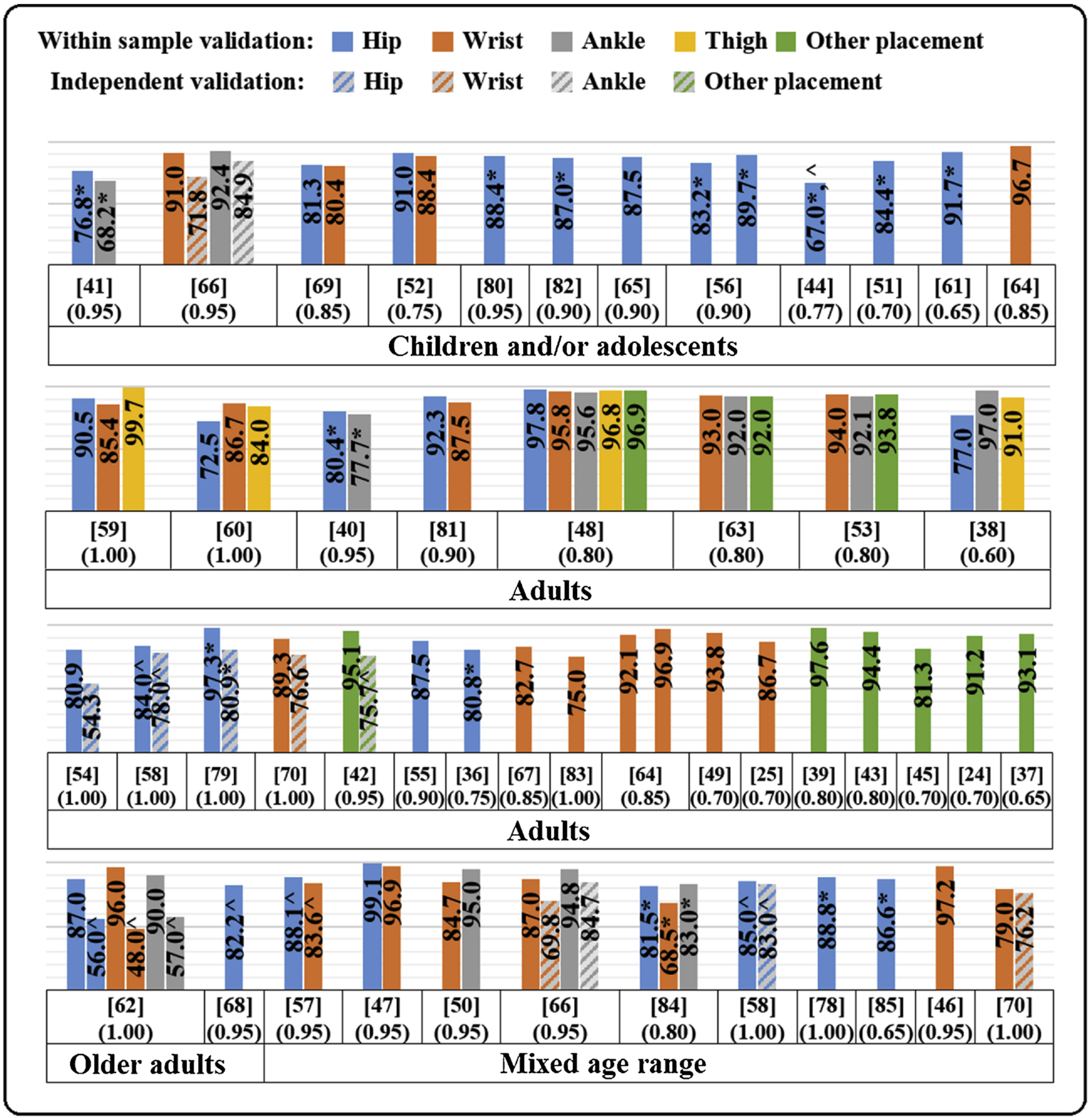

Fig. 2. The extracted prediction accuracies (overall accuracy in %, shown inside the bars) of activity recognition models in relation to accelerometer placement, categorized by the age range of the population using which the models were developed. The literature reference numbers and quality scores are presented in square brackets [] and parentheses (), respectively. *Indicates the data type was activity counts. ˆIndicates the data were acquired in free-living settings.

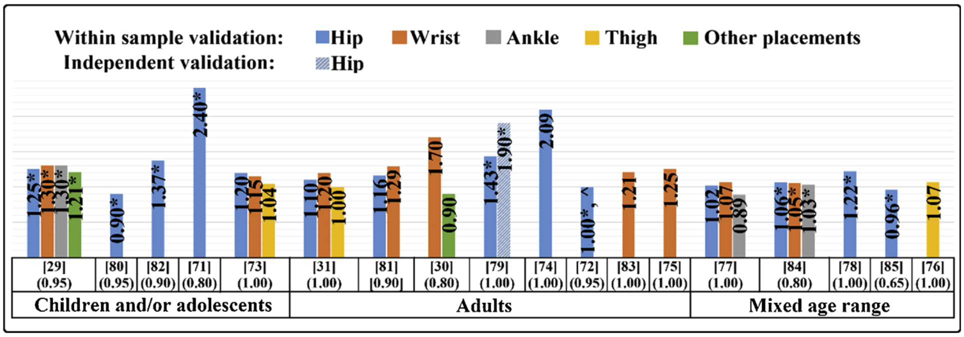

Fig. 3. The extracted prediction accuracies for activity energy expenditure (root mean squared error in MET shown inside the bars) models in relation to accelerometer placement, categorized by the age range of the population using which the models were developed. The literature reference numbers and quality scores are presented square brackets [] and parentheses (), respectively. *Indicates the data type was activity counts. ˆIndicates the data were acquired in free-living settings.

3.4. Data type (raw acceleration data versus activity counts)

3.4.1. Activity recognition

Thirteen studies (27%) calibrated and validated AR models with AC data [36,40,41,44,51,56,61,78–80,82,84,85]. Among these studies, three pointed out that raw acceleration data may be useful for improving the predictive accuracy [40,41,78]. Regarding accelerometer placement, only three of the studies with AC [40,41,84] considered other wear locations than hip (Fig. 2).

3.4.2. Activity energy expenditure

Eleven studies (50%) calibrated and validated AEE with AC data [29,32,33,71,72,78–80,82,84,85]. Among these studies, four pointed out that raw acceleration data may be useful for improving the predictive accuracy [29,72,78,80]. Regarding accelerometer placement, only two of the studies with AC [29,84] considered other wear locations than hip (Fig. 3).

3.5. Sampling frequency

3.5.1. Activity recognition

Thirty-five studies (72%) calibrated and validated AR with raw acceleration data. Across these studies, the sampling frequency varied from 20 to 100 Hz. Seventeen studies developed the models using data sampled at 20 to 30 Hz [25,37,39,42,43,46,52,54,57–60,64,67,68,70,81], eleven studies using data sampled at 50 to 90 Hz [26–28,38,47–50,61,62,83], and seven studies using data sampled at 100 Hz [24,45,53,55,63,64,69].

One study, in which the accelerometer was placed at the wrist, investigated the optimal sampling frequency rate for development of AR models [46]. According to this study, a sampling frequency between 20–25 Hz is enough for developing wrist-based AR models and increasing the sampling frequency above 25 Hz does not improve the predictive accuracy of the models.

3.5.2. Activity energy expenditure

Eleven studies (50%) calibrated and validated AEE with raw acceleration data. Across these studies, the sampling frequency varied from 20 to 100 Hz. Among these studies, seven developed the models using data sampled at 20 to 40 Hz [31,35,73–76,81], three studies using data sampled at 60 to 80 Hz [30,77,83], and one study using data sampled at 100 Hz [27].

3.6. Windowing approach and window size

3.6.1. Activity recognition

The window size varied from 1 to 60 s in the studies that calibrated and validated AR models. Seven studies (15%) used fixed-length overlapping sliding windows with a window size ranging from 2 to 10.24 s and overlapping size ranging from 20% to 75% [24,38,43,48,53,63,64]. Other studies (85%) used fixed-length non-overlapping sliding windows with a window size ranging from 0.5 to 60 s. Out of the studies using non-overlapping windows, seven studies with AC [51,56,61,78–80,82] and four studies with raw acceleration data [57,58,68,81] selected the window size of 60 s. Four studies with raw acceleration selected the window size of 20 or 30 s [49,59,62,70]. Seven studies, one of which with AC [36], selected the window size of 12.8 or 15 s [36,46,47,50,66,69,83]. Eleven studies, three of which with AC [40,41,84], selected the window size from 5 to 10 s [25,34,37,40–42,52,54,60,67,84]. Four studies with raw acceleration data selected the window size from 1 to 2.5 s [26,28,55,65]. One study used AC directly without segmentation [44], one study used varying window length (consisted of an integration and threshold technique) [45], and one study did not report the window size [85].

With respect to age group, across the studies with children and/or adolescents (n = 12), five studies selected non-overlapping window size of 60 s [51,56,61,80,82], one study 12.8 s [66], and one study 15 s [69]. One study used the window size of 10.24 s with 50% overlapping [64]. Three studies used the non-overlapping window size of shorter than 10 s [41,52,65], and one study AC without segmentation [44]. Accordingly, out of the studies with adults (n = 25), two studies selected non-overlapping window size of 60 s [79,81], two studies 30 s [59,70], one study 20 s [49], two studies 15 s [36,83], and one study 12.8 s [25,26,39,40,42,54,55,60,67]. One study used varying window length [45], and seven studies used a window size of 2–10.24 sec with 20%–75% overlapping [24,38,43,48,53,63,64]. Of the three studies with older adults, one study used non-overlapping window size of 60 s [68], one study 30 s [62], and one study 1 s [28].

There was an agreement with the findings between the studies testing the effects of window size on the predictive accuracy of AR models. Two studies with AC in children and/or adolescents reported minimal accuracy reduction when the window size decreased from 60 to 30, 20, 15, and 10 s [56,80], and one study with raw acceleration data in older adults when the window size decreased from 30 to 15, 10, and 5 s [62]. Three studies with raw acceleration data, one of which with children and/or adolescents [50], one with adults [37], and one with mixed age range [66] also reported minimal accuracy reduction when decreased from 12.8 to 8, 6.4, 4, 3.2, 2 and/or 1.6 s. One study with raw acceleration data from adults found 1.5 s as the most optimal window size, while with 1, 2, 2.5, and 3 s windows the results were comparable [55].

3.6.2. Activity energy expenditure

Nineteen studies (86%) calibrated and validated AEE models using fixed-length non-overlapping sliding windows with a window size ranging from 4 to 60 s. The other three studies (14%) used adaptive windowing on the basis of variation in AC [32,33,72]. Out of the studies using non-overlapping windows, five studies with raw acceleration data [71,78–80,82] and four studies with AC selected the window size of 60 s [34,35,76,81]. Six studies with raw acceleration data selected the window size of 30 s [31,73–77]. Three studies, two of which with AC [29,84], selected the window size from 10 to 15 s [29,83,84]. One study with raw acceleration data selected the window size of 4 s [30], and one study did not report the window size [85].

With respect to age group, across the studies with children and/or adolescents (n = 4), three studies selected non-overlapping window size of 60 s [71,80,82], and one study 15 s [29]. Out of the studies with adults (n = 12), three studies selected non-overlapping window size of 60 s [73,79,81], four studies 30 s [31,74–76], one study 15 s [83], and one study 4 s [30]. Two studies used varying window length [32,72].

One study with AC in children and/or adolescents tested the effects of different window size on the predictive accuracy of AEE models and reported minimal accuracy reduction when the window size decreased from 60 to 30, 20, 15, and 10 s [80].

3.7. Feature generation and selection

3.7.1. Activity recognition

Ten studies (20%) used an automated feature selection method in prior to develop AR models [26,34,40,47,53,54,63,64,66] and four studies (8%) tested different manually-provided feature sets [50,60,70,85]. Twenty two studies (46%) with raw acceleration data used a combination of TD and FD features as inputs [25,26,34,37,38,42,45,46,48–50,54,57,58,62,64–69]. Twenty studies (41%), eight of which with AC [36,40,41,56,78–80,82], used only TD features as inputs [36,39–41,43,51–53,55,56,59,60,63,70,78–80,82,83,85]. One study used a combination of TD and wavelet features extracted from AC [61], one study a combination of TD, FD, and wavelet features extracted from raw acceleration data [47], and one study only FD features extracted from raw acceleration data [24]. One study used Ngram features extracted from AC (an approach used in natural language processing) [84], one study used raw acceleration without feature extraction (distance-based method) [28], and one study used AC without feature extraction [44].

There was an agreement with the findings between the studies testing the inclusion of wavelet or distance-based features as additional input features to the combination of TD and/or FD features. Two studies with raw acceleration data reported that the inclusion of wavelet features to a combination of TD and FD features [38,50], and one study with AC reported that the inclusion of distance-based features to TD features [85] minimally improved the predictive accuracy of AR models. Similarly, there was an agreement with the findings between the studies testing the inclusion of FD to TD features when the models were developed within the sample population. Four studies with raw acceleration data reported that inclusion of FD features to TD features improved predictive accuracy of AR models when the models were developed within a population [50,57,66,70]. On the contrary, one study reported that the inclusion of FD features did not improve the predictive accuracy of models [70], whereas another study reported that their inclusion improved the predictive accuracy of models when cross-validated in an independent population [66].

3.7.2. Activity energy expenditure

Two studies (9%) used an automated feature selection method in prior to developing their AEE models [27,30] and three studies (14%) tested different manually-provided feature sets [73,75,85]. Sixteen studies (73%), eight of which with AC [29,32,33,72,78–80,82], used only TD features as inputs [29,31–33,72,79,80,82,83,85]. Four studies (18%) with raw acceleration data used a combination of TD and FD features as inputs [27,30,35,81]. One study used Ngram features extracted from AC [84] and one used AC directly without feature extraction [71].

Two studies with AC tested the inclusion of different input features to TD features, and showed a minimal benefit in the inclusion of distance-based [85] and anthropometrical features [73] to TD features.

3.8. Axes of measurements

3.8.1. Activity recognition

Twenty one studies (43%) that calibrated and validated AR models used all three axes of measurement (i.e., x-, y-, and z-direction) [28,34,37–39,41,42,48,51,53,55,60,62–65,67,70,71,74,82], eleven studies (23%) used the resultant orientation-independent vector magnitude (VM) [25,49,50,57,58,66,68,69,83–85], seven studies (15%) used vertical axis [36,40,46,56,78–80], and four studies (8%) combined all three axes of measurements and VM [26,43,54,81] for developing the models. Two studies used z-axis [24,45], one study a biaxial accelerometer [35], and two studies did not report which axes of measurement used for developing the models [52,61].

There was a somewhat contradictory finding between the studies testing the inclusion of different axes of measurements. One study with raw acceleration data found no benefit in the inclusion of more than one axis for the prediction accuracy of wrist-based AR models [46]. On the contrary, one study with AC found that triaxial data increased the accuracy of both hip- and ankle-based AR models in comparison to uniaxial data [41]. However, the data type was different between the studies.

3.8.2. Activity energy expenditure

Thirteen studies (59%) that calibrated and validated AEE models used all three axes of measurement [27,29–33,71–74,76,77,82] and three studies (14%) used VM for developing the models [83–85]. One study used vertical axis [79] and one study a biaxial accelerometer [35] for developing the models. One study with AC reported that using triaxial data was beneficial compared to vertical axis only for predictive accuracy of hip-based models [72].

3.9. Machine learning technique

3.9.1. Activity recognition

Nineteen studies performed AR with more than one ML technique. There was a consistency between the studies testing different techniques. Six studies reported that the predictive accuracy of different techniques, including ANN, SVM, RF, NB, K-nearest neighbors, and/or DTs were comparable and satisfactory, and therefore they did not specify the algorithms with the highest accuracy [34,46,55,69,83,85]. This was further supported by six studies that also found comparable predictive accuracy with different techniques while the most optimal technique(s) were SVM (compared to ANN, DT, and/or NB) [42,48], SVM and DT (compared to ANN, LR, and NB) [47], SVM and majority voting (compared to ANN and DT) [42], deep-learning neural network (compared to a standard ANN) [56], and rotation forest (compared to ANN and KNN) [53]. On the contrary, there were six studies that found a significantly higher accuracy for ANN (compared to DT, LR, and NB) [25], a Bayes approach (compared to SVM) [45], HMM (compared to quadratic discriminant analysis) [36], RF and majority voting (compared to ANN, DTs, KNN, SVM, and/or majority voting) [64], RF (comparing to ANN, DT, SVM, and majority voting) [70], and a distance-based method (compared to ANN, KNN, citation KNN, and SVM) [82].

3.9.2. Activity energy expenditure

Three studies performed AEE with more than one ML technique. Two of these studies reported comparable accuracy for different techniques, including ANN, DTs, KNN, LR, NB, RF, and/or SVM [83,85]. One study found a significantly higher accuracy for a distance-based method (compared to ANN, KNN, citation KNN, SVM) [82].

3.10. Accelerometer placement

3.10.1. Hip, wrist, ankle, and thigh

3.10.1.1. Activity recognition. Ten studies compared AR models with hip and wrist, six studies with hip and ankle, four studies with hip and thigh, seven studies with wrist and ankle, three studies with wrist and thigh, and two studies with ankle and thigh (Fig. 2). Eight of the studies comparing hip- and wrist-based models reported a higher accuracy for the hip-based model [47,48,52,57,59,69,81,84] and one study reported a higher accuracy for the wrist-based model [60]. One study reported a higher accuracy for the wrist-based model when the models were developed with laboratory-acquired data and a higher accuracy for the hip-based model when the models were developed with free-living data [62]. On average, the absolute difference between accuracy of the hip- and wrist-based models was 5.5% (standard deviation 4.6%). Three of the studies comparing hip- and ankle-based models reported a higher accuracy for the hip-based model [40,41,48] and two studies reported a higher accuracy for the ankle-based model [38,84]. One study reported a higher accuracy for the hip-based model when the models were developed with laboratory-acquired data and a slightly higher accuracy for the ankle-based when the models were developed with free-living data [62]. On average, the absolute difference between accuracy of the hip- and ankle-based models was 5.6% (SD 6.3%). One of the studies comparing hip- and thigh-based models reported a higher accuracy for the hip-based model [48], while three studies reported a higher accuracy for the thigh-based model [38,59,85]. On average, the absolute difference between accuracy of the hip- and thigh-based models was 8.9% (SD 5.6%).

Of the studies comparing wrist- and ankle-based models, three studies reported a higher accuracy for the wrist-based model [48,53,63] and three studies reported a higher accuracy for the ankle-based model [50,66,84]. One study reported a higher accuracy for the wrist-based model when the models were developed with laboratory-acquired data and a higher accuracy for the ankle-based model when the models were developed with free-living data [62]. On average, the absolute difference between accuracy of the wrist- and ankle-based models was 7.1% (SD 5.3%). One of the studies comparing wrist- and thigh-based models reported a higher accuracy for the wrist-based model [60], while two studies reported a higher accuracy for the thigh-based model [48,59]. On average, the absolute difference between accuracy of the wrist- and thigh-based models was 6.0% (SD 7.2%).

One of the studies comparing ankle- and thigh-based models reported a higher accuracy for the ankle-based model by 6% [38] and one study reported a higher accuracy for the thigh-based model by 1% [48].

3.10.1.2. Activity energy expenditure. Six studies compared AEE models with hip and wrist, three studies with hip and ankle, two studies with hip and thigh, three studies with wrist and ankle, and two studies with wrist and thigh, see Fig. 3. Of the studies comparing hip- and wrist-based models, four studies reported a higher accuracy for the hip-based model [29,31,77,81], while two studies reported a higher accuracy for the wrist-based model [73,84]. On average, the absolute difference between accuracy of the hip- and wrist-based models was 0.06 MET (standard deviation 0.04 MET). One of the studies comparing hip- and ankle-based models reported a slightly higher accuracy for the hip-based model [29], while two studies reported a higher accuracy for the ankle-based model [77,84]. On average, the absolute difference between accuracy of the hip- and ankle-based models was 0.73 MET (SD 0.80 MET). Two of the studies comparing hip- and thigh-based models reported a higher accuracy for the thigh-based model by 0.16 MET [73] and 0.10 MET [31].

Two of the studies comparing wrist- and ankle-based models reported a higher accuracy for the ankle-based model [77,84] and one study reported similar accuracies for the ankle- and wrist-based models [29]. On average, the absolute difference between accuracy of the wrist- and ankle-based models was 0.08 MET (SD 0.08 MET). The two studies comparing wrist- and thigh-based models reported a higher accuracy for the thigh-based model by 0.11 MET [73] and 0.20 MET [31].

3.10.2. Other accelerometer placements than hip, wrist, ankle, and thigh

3.10.2.1. Activity recognition. Three studies compared AR models with other placement(s) and hip, wrist, ankle, or thigh (Fig. 2). The other placements used in these three studies were chest [53,63], and chest and lower back (the accuracy of chest-based model was higher) [48]. One study comparing hip-based model and other placement-based model reported a higher accuracy for the hip-based model by 1% [48]. Of the studies comparing wrist- and other placement-based models, two studies reported a higher accuracy for the wrist-based model [53,63] and one study reported a higher accuracy for the other placement-based model [48]. On average, the absolute difference between accuracy of the wrist- and other placement-based models was 0.7% (SD 0.4%). Of the studies comparing ankle- and other placement-based models, two studies reported a higher accuracy for the other placement-based model [48,53] and one study reported similar accuracies for the ankle- and other placement-based models [63]. On average, the absolute difference between accuracy of the ankle- and other placement-based models was 1.0% (SD 0.8%). One study comparing thigh- and other placement-based models reported similar accuracy for the thigh- and other placement-based models [48].

3.10.2.2. Activity energy expenditure. Two studies compared AEE models with other placement(s) and hip, wrist, or ankle (Fig. 3). The other placements used in these two studies were chest [30], and chest and knees (the accuracy of chest-based model was higher) [29]. One study comparing hip- and other placement-based models reported a higher accuracy for the other placement-based model by 0.04 MET [29]. Two studies comparing wrist- and other placement-based models reported a higher accuracy for the other placement-based models by 0.09 MET [29] and 0.70 MET [30]. One study comparing ankle- and other placement-based models reported a higher accuracy for the other placements-based model by 0.09 MET [29].

3.10.3. Generalization capability

3.10.3.1. Activity recognition. Six studies tested the validity of AR models in an independent population, see Fig. 2 [42,54,58,66,70,79]. Two of these studies found accuracy degradation in the heterogeneity due to population age range, as the data acquisition protocols for populations used for model development and independent validation were similar (i.e., in a controlled manner outside and/or inside laboratory). These two studies reported 3–17% accuracy degradation when the models cross-validated on an independent population with different age range [66,70]. Additionally, one study showed 17% accuracy degradation even though population characteristics and data acquisition protocols were similar in both populations [79]. On the other hand, two studies found the main reason for accuracy degradation in the difference between laboratory- and (semi-)free-living-acquired acceleration data and reported 20–30% accuracy reduction when cross-validating the laboratory-calibrated model on data (semi)-free-living settings, even though the population characteristics used for model development and independent validation were similar [42,54]. Additionally, one study reported 2–6% accuracy degradation when free-living-calibrated models were cross-validated in free-living settings, even though the population characteristics were markedly different [58].

3.10.3.2. Activity energy expenditure. Three studies tested the validity of AEE models in an independent population, see Fig. 3 (only one of them is shown) [32,33,79]. One of these studies examined the validity of model in relation to direct observation for assessment of SB in free-living settings [32] and two studies in relation to indirect calorimetry in controlled settings [33,79]. One of the two studies that examined the validity of AEE model in relation to indirect calorimetry reported 30% increase in rMSE when the models were cross-validated in an independent sample [79]. The other study reported the mean absolute percentage error ranging from 17% to 22%, depending on the performed activity [33].

4. Discussion

In this extensive systematic review of 62 published studies, we reviewed the effects of various parameters on the predictive ability of ML-based models as well as the generalization capability of ML-based models to independent populations in free-living settings. Given the heterogeneous nature of studies as well as the abundance of parameters and methodological decisions that affect the results of ML-based models, it is difficult to provide conclusive evidence for the most optimal parameters. Nevertheless, it appears that various ML modeling approaches together with raw acceleration data sampled at 20–30 Hz have provided comparable predictive accuracies regardless of accelerometer placement for both AR and AEE -based models. However, the generalization capability of ML-based models to independent populations in free-living settings is still a challenge and further studies are needed.

4.1. Data type (raw acceleration data versus activity counts)

ML-based models have been developed with both activity counts and raw acceleration data. A considerable number of the studies that used AC for developing the models have pointed out that using raw acceleration data might be beneficial for improving the predictive accuracy of ML-based models. This finding is in agreement with the existing literature suggesting the use of raw acceleration in calibration and validation studies [2,14,86].

Regarding accelerometer placement, it appeared that when raw acceleration data was used, alternative wear locations in addition to hip including wrist, ankle, and thigh were also considered. Several of the included studies reported that AC during different types of activities could be very similar (e.g., sitting and standing) [40,41,58,78]. The transformation of raw data into AC eliminates significant information that can help to distinguish between various activities [40,41,78]. This can affect the ability of ML-based models to differentiate between different activities, due to the similarity of input features to ML-based methods derived from AC time-series for different activities [72]. Therefore, it seems that raw accelerometry together with advanced ML techniques allow for alternative sensor placements in addition to conventional hip-worn accelerometers for both AR and AEE.

4.2. Sampling frequency

Currently, accelerometers are able to measure and save high frequency acceleration data [12,14]. However, recording data at high frequencies might not be practical for long term monitoring due to memory and/or battery exhaustion [6,46]. Even though only one study has demonstrated that using data sampled at higher frequency than 25 Hz has no benefits for predictive accuracy of wrist-based AR models [46], it appears that sampling frequency above 30 Hz might not be needed for development of ML-based models. Indeed, a majority of studies have used data sampled at 20–30 Hz.

However, it is not yet possible to confirm whether the need of relatively low sampling frequency data is a unique advantage of ML-based approaches over other traditional statistical analytics. This is mainly because raw acceleration data is relatively new and calibration studies are still being developed [3]. A recent systematic review on the calibration of raw accelerometer data partially supports that statistical approaches tend to use higher sampling frequencies between 60–100 Hz, whereas ML-based modeling approaches used relatively lower sampling frequency [6]. This partially supports that ML modeling approaches might offer the opportunity of calibration and validation of accelerometer-based activity monitors using data sampled at relatively lower sampling frequency.

4.3. Windowing approach and window size

The most widely used segmentation method was the fixed-size sliding window approach. Window sizes ranged from 1 to 60 s in the included studies, with some studies including a degree of overlap between windows. However, it seems that in response to the recommendation for using shorter window size [5,9,87,88], a majority of AR studies used relatively short window size between 5 to 15 s for AR. On the contrary, it appeared that a longer window length was used with AC data. This adds to the growing body of evidence that ML-based modeling approaches together with raw accelerometry provide several potential opportunities including AR using relatively shorter window size. Notably, with respect to age group, studies with children and/or adolescents had mostly selected the window size of 30 s or higher. Additionally, different trend was seen for ML-based AEE models, selecting a longer window size. According to the existing literature, it would be preferable to limit window size to smaller values, preferably between 5 to 15 s in order to better describe activities and estimate AEE in free-living settings [5,9,87,88]. This is even more important in children and adolescents since their physical activity patterns are more sporadic [2,6,13]. Future studies focusing on estimation of AEE in children, and/or adolescents should consider shorter window size to effectively monitor more sporadic physical activity patterns [2,6,13].

Smaller window size can lead to accuracy loss, but notably, the accuracy degradation was minimal and the accuracies remained still acceptable in the included studies. This was observed in different age groups. This is promising, as it suggests that ML-based modeling approaches can provide comparable results with minimal accuracy reduction using relatively shorter window size, as opposed to cut-point methods where the results vary greatly depending on the window size [89]. However, the issue of transitive activities has been almost completely neglected in the field. All but three of the included studies assumed that each window interval consists of only a single activity. The transition from one activity type to another under free-living conditions can be problematic for ML-based modeling approaches, since they rely on the extracted features within a predetermined period [2,62,72]. Further research in free-living settings using criterion measures (e.g., direct observation) is needed to elucidate the accuracy of different window sizes. To be better suited for free-living applications, perhaps, future studies will need to develop models that use transition points to define window boundaries rather than sliding windows [2,13,62,72], which could be more feasible with the presence of raw data [2,72].

4.4. Feature generation and selection, axes of measurements

Input features is one of the most primary factors, if is not the most important factor, that directly affects the results of ML-based models [90]. AC-based studies demonstrated that extracting TD features from all three axes of measurement is beneficial for the predictive accuracy of ML-based models. However, contradictory results from raw acceleration-based studies (only one study) suggests that the inclusion of more than one axis of measurement is not beneficial for the predictive accuracy of wrist-based AR models. Extraction of FD features in addition to TD features is prevailing with raw accelerometry. Controversy exists across the included studies regarding the importance of FD features. The contradictions are mainly explained by the fact that input features may need to vary according to the prediction goal and accelerometer placement [2,13,57], while the studies rarely applied an automated feature selection procedure to find the optimal features in relation to accelerometer placement. As the tested feature sets in some studies were manually-provided, an optimized combination of FD and TD features selected by automated feature selection methods might resolve the contradictions.

Therefore, it remains still open which features from which axes of measurement should be used in ML-based methods when developing AR and AEE models in accordance to accelerometer placement. Similarly, it is not clear whether extracting features from 3D raw acceleration data improves upon orientation-independent VM. A comparative study on the effects of different feature sets, selected by automated feature selection methods, on the results of ML-based models is still lacking. Future studies should consider running different feature selection procedures independently for various accelerometer placements in order to find optimal features in accordance to both accelerometer placement and prediction goals. These location-optimized feature sets can potentially lead to improvement and less variation in the results of ML-based models [2,57]. To find location-optimized feature sets, a wide set of combination of TD and FD features extracted from all the axes of measurement and VM (and possibly demographic and anthropometric features) should be fed into feature selection procedures [2,13].

4.5. Machine learning technique

It might be difficult to find an ML technique that is universally better than others for either AR or AEE [85]. This is mainly because the predictive accuracy of ML techniques can vary depending on the feature sets, wear location, and even PA and SB constructs being assessed [91]. ANN, SVM, and RF were the most commonly applied techniques among the included studies. No agreement was observed between the included studies on which ML technique was the most feasible technique. This has been also observed in the existing literature [6]. A majority of studies comparing ML techniques on the AR and AEE tasks consistently found that the predictive accuracy of different ML techniques are comparable, suggesting that certain ML techniques could be adopted in order to harmonize data processing methods, and avoid proliferation of data processing methods and methodological discrepancies [9,14].

Future studies should consider applying various ML techniques in order to enable comparison between different techniques and find the optimal methods in relation to prediction goals, accelerometer placement, and population group.

4.6. Accelerometer placement

Accelerometer placement is one of the main cause of incomparability of the results across the PA and SB studies (e.g., observational, intervention, etc.) [5]. To date, there has been no consensus on a single accelerometer placement within the literature [92]. Perhaps the reason why no consensus has been achieved yet can be explained by the high variation in accuracy of earlier modeling approaches depending on accelerometer placement [93,94]. Similarly, there appears to be no agreement on a single optimal accelerometer placement which would provide the highest accuracy. However, notably, accuracies of ML-based models can vary (for both AR and AEE estimation tasks), but not greatly depending on the sensor placement. Studies comparing different ML-based models with data from various accelerometer placements have clearly stated that the accuracy of models are close to each other and acceptable [29,47,48,57,60,69,77,81], except a few studies in which the difference was substantial which can be explained by study design [34,50,66]. Probably, a consensus for a standardized location might be achieved, considering that advanced ML modeling approaches may provide the opportunity of processing accelerometer data with minimal accuracy loss according to the wear location.

Due to the high compliance and feasibility of wrist-worn accelerometers, wrist was the placement of choice in several large studies including the National Health and Nutrition Examination Survey (NHANES) and the UK Biobank study [7,12]. It seems very realistic that wrist would be the preferred choice of placement in a majority of future studies. This signifies the finding that ML approaches have minimized the differences in predictive accuracies due to wear location [93,94]. Given that ML approaches have enabled precise measurement of SB and PA from wrist-based data comparable to other wear locations [47], it might be possible to adopt wrist as the common wear location and enable the comparability of accelerometry results across the studies. However, over longer period, the minimal differences in results in relation to accelerometer placement could change to substantially different results [29,95]. Additional independent validation studies under free-living conditions are still needed to further confirm whether the minimal variability of results according to accelerometer placement is transferable to free-living settings.

4.7. Generalization capability

ML-based models were rarely validated independently. It is widely accepted that LOSO-based validation provides a good estimate of how ML-based models will be generalized [23]. Arguably, four of the included studies well discussed and showed that LOSO-based validation overestimated the accuracy of ML-based AR models [26,54,62,70]. This issue has also been shown for ML-based AEE models [79]. This is implying the development and evaluation of the method on the same subjects (LOSO or random split) might encourage the ML methods to overfit the data, resulting evaluation statistics validation to be biased and overly optimistic.

When ML-based models are cross-validated on an independent population, unseen activities which were not part of model development along with heterogeneity in sample characteristics are among the other important reasons of accuracy reduction [54,79]. More importantly, the accuracy reduction becomes more severe when laboratory-calibrated models are cross-validated to (semi-)free-living settings [42,54,62]. It was observed that when AR models were cross-validated on an independent population, the overall accuracy often remained higher than the accepted level of 80% suggested in previous studies [62]. However, when the data of independent population was acquired under free-living, the overall accuracies remained below the accepted level of 80% (except one method). This may be because the duration of activities (i.e., transitive activities), types, and acceleration signals (both AC and raw acceleration data) vary greatly in free-living settings compared to laboratory-controlled conditions [58,72]. It appears that the high accuracy of laboratory-calibrated models will not be translated to free-living settings and accuracy loss could be substantial and often not clinically acceptable [26,42,58,62]. Additionally, LOSO-based validation (as well as random split) alone might not be enough for understanding that how ML-based models will perform in a new population, even if the population characteristics remain homogenous [26,54,62,70]. Reporting the predictive accuracy of ML methods for each subject is needed to understand how inter-subject variability will affect the performance of ML models. Future studies should develop ML models using data from more subjects and activity types in order to improve the predictive and generalization capability. In line with previous research, demanding independent validation in free-living settings [6,96], future studies should also move beyond LOSO-based validation towards independent validation of ML-based models in free-living settings on both homogenous and heterogeneous populations.

4.8. Study limitation

The current study has some limitations. Using data fusion of multiple sensors remained outside the scope of this review. Furthermore, we did not focus on the predictive accuracy of ML-based modeling approaches in relation to activity types (e.g., household activities) together with accelerometer placement. If measuring types of activities is of interest, a specific accelerometer placement might provide substantially better results. However, the focus was on the overall predictive ability of ML-based models, as in many applications (e.g., surveillance studies) a wide range of activities including both SB and PA under free-living conditions is of interest. Comparing the predictive accuracy of ML modeling approaches to regression- and cut-point-based methods remained outside the scope of this review. Instead, we assumed that ML models are superior to the traditional statistical procedures.

5. Conclusions

In summary, based on the included papers, our review highlights that relatively low frequency raw acceleration data together with various ML-based modeling approaches offers opportunities for predicting activity type, intensity and energy expenditure with comparable overall predictive accuracies regardless of accelerometer placement. These opportunities also include minimizing accuracy variation due to window size. However, raw accelerometry has also led to methodological discrepancies, since it is not clear which features from which signals should be extracted. Furthermore, most studies relied on laboratory-controlled data, fixed-length non-overlapping windows, and LOSO validation for calibration and validation of accelerometer-based activity monitors using ML-based methods. Due to these reasons, it appeared that the high predictive accuracy of laboratory-calibrated models has not been reproducible in free-living settings. Moving towards independent validation of ML-based models under free-living settings is required in order to develop models with clinically acceptable accuracy in free-living settings and independent populations.

6. Recommandations for future studies and directions

In light of the current existing knowledge and knowledge gaps, for addressing the gaps as a move towards consensus and standardization, future studies should consider the following key points:

- Raw acceleration data allow for AR and AEE estimation from various wear locations including the conventional hip placement as well as alternative placements such as wrist.

- A sampling frequency of 20–30 Hz could be sufficient for both tasks of AEE and AR.
- ML-based modeling approaches are able to classify activities using a short window size of 5–15 sec.
- Input features should be selected in relation to accelerometer placement and prediction task (i.e., AEE or AR) from a broad list of FD and TD features extracted from different axes of measurement (and possibly demographic and anthropometric features) using automatic features selection procedures rather than manually-provided feature sets.
- ANN, SVM, and RF are among the most commonly employed ML-based algorithms, which can be proper candidates for developing of ML-based models.
- Development of ML-based models using free-living- or semi-free-living-acquired data rather than laboratory-acquired data.
- Moving beyond LOSO-based validation towards independent validation of ML-based models in free-living settings on both homogenous and heterogeneous populations.
- With the presence of raw acceleration data, developing methods for identifying the transition points between activities rather than relying on fixed-length windows.

Conflict of interest

The authors attest that they have no conflicts of interest to disclose. This paper reflects the viewpoints of the study authors only.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 713645, the Ministry of Education and Culture in Finland [grant numbers OKM/86/626/2014, OKM/43/626/2015, OKM/17/626/2016], Infotech Oulu, Finland, and Northern Ostrobothnia Hospital District.

References

[1] C.A. Celis-Morales, F. Perez-Bravo, L. Ibanez, C. Salas, M.E.S. Bailey, J.M.R. Gill, Objective vs. Self-reported physical activity and sedentary time: effects of measurement method on relationships with risk biomarkers, PLoS One 7 (2012) e36345.

[2] D.R. Bassett Jr, A.V. Rowlands, S.G. Trost, Calibration and validation of wearable monitors, Med. Sci. Sports Exerc. 44 (2012) S32–S38.

[3] P. Freedson, H.R. Bowles, R. Troiano, W. Haskell, Assessment of physical activity using wearable monitors: recommendations for monitor calibration and use in the field, Med. Sci. Sports Exerc. 44 (2012) S1–S4.

[4] I.-M. Lee, E.J. Shiroma, Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges, Br. J. Sport. Med. 48 (2014) 197–201.

[5] S.J. Strath, K.A. Pfeiffer, M.C. Whitt-Glover, Accelerometer use with children, older adults, and adults with functional limitations, Med. Sci. Sports Exerc. 44 (2012) S77–S85.

[6] M. de Almeida Mendes, I.C.M. da Silva, V.V. Ramires, F.F. Reichert, R.C. Martins, E. Tomasi, Calibration of raw accelerometer data to measure physical activity: a systematic review, Gait Posture 61 (2018) 98–110.

[7] R.P. Troiano, J.J. McClain, R.J. Brychta, K.Y. Chen, Evolution of accelerometer methods for physical activity research, Br. J. Sport. Med. 48 (2014) 1019–1023.

[8] B.H. Hansen, E. Kolle, S.M. Dyrstad, I. Holme, S.A. Anderssen, Accelerometer-determined physical activity in adults and older people, Med. Sci. Sports Exerc. 44 (2012) 266–272.

[9] K.L. Cain, J.F. Sallis, T.L. Conway, D. Van Dyck, L. Calhoon, Using accelerometers in youth physical activity studies: a review of methods, J. Phys. Act. Heal. 10 (2013) 437–450.

[10] Y. Kim, M.W. Beets, G.J. Welk, Everything you wanted to know about selecting the "right" Actigraph accelerometer cut-points for youth, but…: a systematic review, J. Sci. Med. Sport 15 (2012) 311–321.

[11] X. Janssen, D.P. Cliff, Issues related to measuring and interpreting objectively measured sedentary behavior data, Meas. Phys. Educ. Exerc. Sci. 19 (2015) 116–124.

[12] V.T. van Hees, K. Thaler-Kall, K.-H. Wolf, J.C. Brønd, A. Bonomi, M. Schulze, et al., Challenges and opportunities for harmonizing research methodology: raw accelerometry, Methods inf, Med. 55 (2016) 525–532.

[13] S. Liu, R. Gao, P. Freedson, Computational methods for estimating energy expenditure in human physical activities, Med. Sci. Sports Exerc. 44 (2012) 2138–2146.

[14] K. Wijndaele, K. Westgate, S.K. Stephens, S.N. Blair, F.C. Bull, S.F.M. Chastin, et al., Utilization and harmonization of adult accelerometry data: review and expert consensus, Med. Sci. Sports Exerc. 47 (2015) 2129–2139.

[15] M. Aittasalo, H. Vähä-Ypyä, T. Vasankari, P. Husu, A.-M. Jussila, H. Sievänen, Mean amplitude deviation calculated from raw acceleration data: a novel method for classifying the intensity of adolescents' physical activity irrespective of accelerometer brand, BMC Sports Sci. Med. Rehabil. 7 (2015) 18.

[16] K. Bakrania, T. Yates, A.V. Rowlands, D.W. Esliger, S. Bunnewell, J. Sanders, et al., Intensity thresholds on raw acceleration data: euclidean norm minus one (ENMO) and mean amplitude deviation (MAD) approaches, PLoS One 11 (2016) e0164045.

[17] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, PRISMA Group, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, PLoS Med. 6 (2009) e1000097.

[18] A.M. Khan, Y.-K. Lee, S. Lee, T.-S. Kim, Accelerometer's position independent physical activity recognition system for long-term activity monitoring in the elderly, Med. Biol. Eng. Comput. 48 (2010) 1271–1279.

[19] U. Maurer, A. Smailagic, D.P. Siewiorek, M. Deisher, Activity Recognition and Monitoring Using Multiple Sensors on Different Body Positions, Int. Work. Wearable Implant. Body Sens. Networks, (2006), pp. 113–116.

[20] L.M. Kmet, R.C. Lee, L.S. Cook, Standard quality assessment criteria for evaluating primary research papers from a variety of fields. HTA Initiative #13, Alberta Herit. Found. Med. Res. (AHFMR), Edmont. (2004).

[21] G.J. Welk, Principles of design and analyses for the calibration of accelerometry-based activity monitors, Med. Sci. Sports Exerc. 37 (2005) S501–S511.

[22] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, et al., Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (2008) 1–37.

[23] J. Staudenmayer, W. Zhu, D.J. Catellier, Statistical considerations in the analysis of accelerometry-based activity monitor data, Med. Sci. Sports Exerc. 44 (2012) S61–S67.

[24] L. Oudre, J. Jakubowicz, P. Bianchi, C. Simon, Classification of periodic activities using the Wasserstein distance, IEEE Trans. Biomed. Eng. 59 (2012) 1610–1619.

[25] J. Margarito, R. Helaoui, A.M. Bianchi, F. Sartor, A.G. Bonomi, User-independent recognition of sports activities from a single wrist-worn accelerometer: a template-matching-based approach, IEEE Trans. Biomed. Eng. 63 (2016) 788–796.

[26] V.T. van Hees, R. Golubic, U. Ekelund, S. Brage, Impact of study design on development and evaluation of an activity type classifier, J. Appl. Physiol. 114 (2013) 1042–1051.

[27] L. Atallah, J.J. Leong, B. Lo, G.-Z. Yang, Energy expenditure prediction using a miniaturized ear-worn sensor, Med. Sci. Sports Exerc. 43 (2011) 1369–1377.

[28] B. He, J. Bai, V.V. Zipunnikov, A. Koster, P. Caserotti, B. Lange-Maia, et al., Predicting human movement with multiple accelerometers using movelets, Med. Sci. Sports Exerc. 46 (2014) 1859–1866.

[29] K.A. Mackintosh, A.H.K. Montoye, K.A. Pfeiffer, M.A. McNarry, Investigating optimal accelerometer placement for energy expenditure prediction in children using a machine learning approach, Physiol. Meas. 37 (2016) 1728–1740.

[30] M. Altini, J. Penders, R. Vullers, O. Amft, Estimating energy expenditure using body-worn accelerometers: a comparison of methods, sensors number and positioning, IEEE J. Biomed. Heal. Informatics. 19 (2015) 219–226.

[31] A.H.K. Montoye, M. Begum, Z. Henning, K.A. Pfeiffer, Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data, Physiol. Meas. 38 (2017) 343–357.

[32] Y. Kim, V.W. Barry, M. Kang, Validation of the ActiGraph GT3X and activPAL accelerometers for the assessment of sedentary behavior, Meas. Phys. Educ. Exerc. Sci. 19 (2015) 125–137.

[33] L.D. Ellingson, I.J. Schwabacher, Y. Kim, G.J. Welk, D.B. Cook, Validity of an integrative method for processing physical activity data, Med. Sci. Sports Exerc. 48 (2016) 1629–1638.

[34] L. Atallah, B. Lo, R. King, G.-Z. Yang, Sensor positioning for activity recognition using wearable accelerometers, IEEE Trans. Biomed. Circuits Syst. 5 (2011) 320–329.

[35] M.P. Rothney, M. Neumann, A. Béziat, K.Y. Chen, An artificial neural network model of energy expenditure using nonintegrated acceleration signals, J. Appl. Physiol. 103 (2007) 1419–1427.

[36] D.M. Pober, J. Staudenmayer, C. Raphael, P.S. Freedson, Development of novel techniques to classify physical activity mode using accelerometers, Med. Sci. Sports Exerc. 38 (2006) 1626–1634.

[37] A.G. Bonomi, A.H.C. Goris, B. Yin, K.R. Westerterp, Detection of type, duration, and intensity of physical activity using an accelerometer, Med. Sci. Sport. Exerc. 41 (2009) 1770–1777.

[38] S.J. Preece, J.Y. Goulermas, L.P.J. Kenney, D. Howard, A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data, IEEE Trans. Biomed. Eng. 56 (2009) 871–879.

[39] A.M. Khan, Y.-K. Lee, S.Y. Lee, T.-S. Kim, A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer, IEEE Trans. Inf. Technol. Biomed. 14 (2010) 1166–1172.

[40] S.I. De Vries, F.G. Garre, L.H. Engbers, V.H. Hildebrandt, S. Van Buuren, Evaluation of neural networks to identify types of activity using accelerometers, Med. Sci. Sport. Exerc. 43 (2011) 101–107.

[41] S.I. De Vries, M. Engels, F.G. Garre, Identification of children's activity type with accelerometer-based neural networks, Med. Sci. Sports Exerc. 43 (2011) 1994–1999.

[42] I.C. Gyllensten, A.G. Bonomi, Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life, IEEE Trans.

Biomed. Eng. 58 (2011) 2656–2663.

[43] M.-W. Lee, A.M. Khan, T.-S. Kim, A single tri-axial accelerometer-based real-time personal life log system capable of human activity recognition and exercise information generation, Pers. Ubiquitous Comput. 15 (2011) 887–898.

[44] N. Ruch, M. Rumo, U. Mäder, Recognition of activities in children by two uniaxial accelerometers in free-living conditions, Eur. J. Appl. Physiol. 111 (2011) 1917–1927.

[45] M. Schmid, F. Riganti-Fulginei, I. Bernabucci, A. Laudani, D. Bibbo, R. Muscillo, et al., SVM versus MAP on accelerometer data to distinguish among locomotor activities executed at different speeds, Comput. Math. Methods Med. (2013) (2013).

[46] S. Zhang, P. Murray, R. Zillmer, R.G. Eston, M. Catt, A.V. Rowlands, Activity classification using the GENEA: optimum sampling frequency and number of axes, Med. Sci. Sport. Exerc. 44 (2012) 2228–2234.

[47] S. Zhang, A.V. Rowlands, P. Murray, T.L. Hurst, Physical activity classification using the GENEA wrist-worn accelerometer, Med. Sci. Sport. Exerc. 44 (2012) 742–748.

[48] I. Cleland, B. Kikhia, C. Nugent, A. Boytsov, J. Hallberg, K. Synnes, et al., Optimal placement of accelerometers for the detection of everyday activities, Sensors 13 (2013) 9183–9200.

[49] D. John, J. Sasaki, J. Staudenmayer, M. Mavilia, P.S. Freedson, Comparison of raw acceleration from the GENEA and ActiGraph^{TM} GT3X+ activity monitors, Sensors 13 (2013) 14754–14763.

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

[51] W. Zhao, A.L. Adolph, M.R. Puyau, F.A. Vohra, N.F. Butte, I.F. Zakeri, Support vector machines classifiers of physical activities in preschoolers, Physiol. Rep. 1 (2013) e00006.

[52] S.G. Trost, Y. Zheng, W.-K. Wong, Machine learning for activity recognition: hip versus wrist data, Physiol. Meas. 35 (2014) 2183–2189.

[53] M. Arif, A. Kattan, Physical activities monitoring using wearable acceleration sensors attached to the body, PLoS One 10 (2015) e0130851.

[54] T. Bastian, A. Maire, J. Dugas, A. Ataya, C. Villars, F. Gris, et al., Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: laboratory-based calibrations are not enough, J. Appl. Physiol. 118 (2015) 716–722.

[55] B. Fida, I. Bernabucci, D. Bibbo, S. Conforto, M. Schmid, Varying behavior of different window sizes on the classification of static and dynamic physical activities from a single accelerometer, Med. Eng. Phys. 37 (2015) 705–711.

[56] M. Hagenbuchner, D.P. Cliff, S.G. Trost, N. Van Tuc, G.E. Peoples, Prediction of activity type in preschool children using machine learning techniques, J. Sci. Med. Sport 18 (2015) 426–431.

[57] K. Ellis, J. Kerr, S. Godbole, J. Staudenmayer, G. Lanckriet, Hip and wrist accelerometer algorithms for free-living behavior classification, Med. Sci. Sports Exerc. 48 (2016) 933–940.

[58] J. Kerr, R.E. Patterson, K. Ellis, S. Godbole, E. Johnson, G. Lanckriet, et al., Objective assessment of physical activity: classifiers for public health, Med. Sci. Sports Exerc. 48 (2016) 951–957.

[59] A.H.K. Montoye, J.M. Pivarnik, L.M. Mudd, S. Biswas, K.A. Pfeiffer, Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior, AIMS Public Heal. 3 (2016) 298–312.

[60] A.H.K. Montoye, J.M. Pivarnik, L.M. Mudd, S. Biswas, 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. Exerc. Sci. 20 (2016) 173–183.

[61] X. Ren, W. Ding, S.E. Crouter, Y. Mu, R. Xie, Activity recognition and intensity estimation in youth from accelerometer data aided by machine learning, Appl. Intell. 45 (2016) 512–529.

[62] J.E. Sasaki, A. Hickey, J. Staudenmayer, D. John, J.A. Kent, P.S. Freedson, Performance of activity classification algorithms in free-living older adults, Med. Sci. Sports Exerc. 48 (2016) 941–950.

[63] M. Arif, A. Kattan, S.I. Ahamed, Classification of physical activities using wearable sensors, Intell. Autom. Soft Comput. 23 (2017) 21–30.

[64] A.K. Chowdhury, D. Tjondronegoro, V. Chandran, S.G. Trost, Ensemble methods for classification of physical activities from wrist accelerometry, Med. Sci. Sports Exerc. 49 (2017) 1965–1973.

[65] J. Kühnhausen, J. Dirk, F. Schmiedek, Individual classification of elementary school children's physical activity: a time-efficient, group-based approach to reference measurements, Behav. Res. Methods 49 (2017) 685–697.

[66] A. Mannini, M. Rosenberger, W.L. Haskell, A.M. Sabatini, S.S. Intille, Activity recognition in youth using single accelerometer placed at wrist or ankle, Med. Sci. Sports Exerc. 49 (2017) 801–812.

[67] T.G. Pavey, N.D. Gilson, S.R. Gomersall, B. Clark, S.G. Trost, Field evaluation of a random forest activity classifier for wrist-worn accelerometer data, J. Sci. Med. Sport 20 (2017) 75–80.

[68] D. Rosenberg, S. Godbole, K. Ellis, C. DI, A. Lacroix, L. Natarajan, et al., Classifiers for accelerometer-measured behaviors in older women, Med. Sci. Sport. Exerc. 49 (2017) 610–616.

[69] S.G. Trost, D. Cliff, M. Ahmadi, N. Van Tuc, M. Hagenbuchner, Sensor-enabled activity class recognition in preschoolers: hip versus wrist data, Med. Sci. Sports Exerc. 50 (2017) 634–641.

[70] A.H.K. Montoye, B.S. Westgate, M.R. Fonley, K.A. Pfeiffer, Cross-validation and out-of-sample testing of physical activity intensity predictions using a wrist-worn accelerometer, J. Appl. Physiol. 124 (2018) 1284–1293.

[71] N. Ruch, F. Joss, G. Jimmy, K. Melzer, J. Hänggi, U. Mäder, Neural network versus activity-specific prediction equations for energy expenditure estimation in children, J. Appl. Physiol. 115 (2013) 1229–1236.

[72] K. Lyden, S.K. Keadle, J. Staudenmayer, P.S. Freedson, A method to estimate free-living active and sedentary behavior from an accelerometer, Med. Sci. Sports Exerc. 46 (2014) 386–397.

[73] A.H. Montoye, L.M. Mudd, S. Biswas, K.A. Pfeiffer, Energy expenditure prediction using raw accelerometer data in simulated free living, Med. Sci. Sport. Exerc. 47 (2015) 1735–1746.

[74] A.H.K. Montoye, B. Dong, S. Biswas, K.A. Pfeiffer, Validation of a wireless accelerometer network for energy expenditure measurement, J. Sports Sci. 34 (2016) 2130–2139.

[75] A.H.K. Montoye, J.M. Pivarnik, L.M. Mudd, S. Biswas, K.A. Pfeiffer, Wrist-independent energy expenditure prediction models from raw accelerometer data, Physiol. Meas. 37 (2016) 1770–1784.

[76] A.H.K. Montoye, J.M. Pivarnik, L.M. Mudd, S. Biswas, K.A. Pfeiffer, Evaluation of the activPAL accelerometer for physical activity and energy expenditure estimation in a semi-structured setting, J. Sci. Med. Sport 20 (2017) 1003–1007.

[77] A.H.K. Montoye, S.A. Conger, C.P. Connolly, M.T. Imboden, M.B. Nelson, J.M. Bock, et al., Validation of accelerometer-based energy expenditure prediction models in structured and simulated free-living settings, Meas. Phys. Educ. Exerc. Sci. 21 (2017) 223–234.

[78] J. Staudenmayer, D. Pober, S. Crouter, D. Bassett, P. Freedson, An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer, J. Appl. Physiol. 107 (2009) 1300–1307.

[79] P.S. Freedson, K. Lyden, S. Kozey-Keadle, 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. 111 (2011) 1804–1812.

[80] S.G. Trost, W.-K. Wong, K.A. Pfeiffer, Y. Zheng, Artificial neural networks to predict activity type and energy expenditure in youth, Med. Sci. Sports Exerc. 44 (2012) 1801–1809.

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

[82] Y. Mu, H.Z. Lo, W. Ding, K. Amaral, S.E. Crouter, Bipart: learning block structure for activity detection, IEEE Trans. Knowl. Data Eng. 26 (2014) 2397–2409.

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

[84] S.J. Strath, R.J. Kate, K.G. Keenan, W.A. Welch, A.M. Swartz, Ngram time series model to predict activity type and energy cost from wrist, hip and ankle accelerometers: implications of age, Physiol. Meas. 36 (2015) 2335–2351.

[85] R.J. Kate, A.M. Swartz, W.A. Welch, S.J. Strath, Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data, Physiol. Meas. 37 (2016) 360–379.

[86] G. Plasqui, Ag. Bonomi, K.R. Westerterp, Daily physical activity assessment with accelerometers: new insights and validation studies, Obes. Rev. 14 (2013) 451–462.

[87] D.P. Heil, S. Brage, M.P. Rothney, Modeling physical activity outcomes from wearable monitors, Med. Sci. Sport. Exerc. 44 (2012) S50–S60.

[88] C.E. Matthews, M. Hagströmer, D.M. Pober, H.R. Bowles, Best practices for using physical activity monitors in population-based research, Med. Sci. Sports Exerc. 44 (2012) S68–S76.

[89] J.A. Banda, K.F. Haydel, T. Davila, M. Desai, S. Bryson, W.L. Haskell, et al., Effects of varying epoch lengths, wear time algorithms, and activity cut-points on estimates of child sedentary behavior and physical activity from accelerometer data, PLoS One 11 (2016) e0150534.

[90] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[91] K. Lyden, S.L. Kozey, J.W. Staudenmeyer, P.S. Freedson, A comprehensive evaluation of commonly used accelerometer energy expenditure and MET prediction equations, Eur. J. Appl. Physiol. 111 (2011) 187–201.

[92] D.R. Bassett, R.P. Troiano, J.J. McClain, D.L. Wolff, Accelerometer-based physical activity: total volume per day and standardized measures, Med. Sci. Sports Exerc. 47 (2015) 833–838.

[93] M.E. Rosenberger, W.L. Haskell, F. Albinali, S. Mota, J. Nawyn, S. Intille, Estimating activity and sedentary behavior from an accelerometer on the hip or wrist, Med. Sci. Sports Exerc. 45 (2013) 964–975.

[94] M.C. Schall Jr, N.B. Fethke, H. Chen, Evaluation of four sensor locations for physical activity assessment, Appl. Ergon. 53 (2016) 103–109.

[95] J. Kerr, C.R. Marinac, K. Ellis, S. Godbole, A. Hipp, K. Glanz, et al., Comparison of accelerometry methods for estimating physical activity, Med. Sci. Sport. Exerc. 49 (2016) 617–624.

[96] S.K. Keadle, Video-recorded direct observation: a step forward for physical activity measurement, Med. Sci. Sport. Exerc. 50 (2018) 1313–1314.