Gait & Posture 90 (2021) 120–128

Contents lists available at ScienceDirect

Gait & Posture

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

Review

Feature selection for unsupervised machine learning of accelerometer data physical activity clusters – A systematic review

Petra J. Jones [a,b,*], Mike Catt [c], Melanie J. Davies [a,b,d], Charlotte L. Edwardson [b,d], Evgeny M. Mirkes [e], Kamlesh Khunti [a,d], Tom Yates [b,d], Alex V. Rowlands [b,d,f]

[a] Leicester Diabetes Centre, University Hospitals of Leicester, Leicester General Hospital, Gwendolen Road, Leicester LE5 4PW, UK
[b] Diabetes Research Centre, University of Leicester, Leicester General Hospital, Gwendolen Road, Leicester LE5 4PW, UK
[c] Population Health Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, UK
[d] NIHR Leicester Biomedical Research Centre, Leicester General Hospital, Gwendolen Road, Leicester, LE5 4PW, UK
[e] School of Mathematics and Actuarial Science, University of Leicester, University Road, Leicester LE1 7RD, UK
[f] Alliance for Research in Exercise, Nutrition and Activity (ARENA), Sansom Institute for Health Research, Division of Health Sciences, University of South Australia, Adelaide, Australia

ARTICLE INFO

Keywords:
Clustering
Physical activity
Accelerometer
Feature selection
K-means

ABSTRACT

Background: Identifying clusters of physical activity (PA) from accelerometer data is important to identify levels of sedentary behaviour and physical activity associated with risks of serious health conditions and time spent engaging in healthy PA. Unsupervised machine learning models can capture PA in everyday free-living activity without the need for labelled data. However, there is scant research addressing the selection of features from accelerometer data. The aim of this systematic review is to summarise feature selection techniques applied in studies concerned with unsupervised machine learning of accelerometer-based device obtained physical activity, and to identify commonly used features identified through these techniques. Feature selection methods can reduce the complexity and computational burden of these models by removing less important features and assist in understanding the relative importance of feature sets and individual features in clustering.
Method: We conducted a systematic search of Pubmed, Medline, Google Scholar, Scopus, Arxiv and Web of Science databases to identify studies published before January 2021 which used feature selection methods to derive PA clusters using unsupervised machine learning models.
Results: A total of 13 studies were eligible for inclusion within the review. The most popular feature selection techniques were Principal Component Analysis (PCA) and correlation-based methods, with k-means frequently used in clustering accelerometer data. Cluster quality evaluation methods were diverse, including both external (e.g. cluster purity) or internal evaluation measures (silhouette score most frequently). Only four of the 13 studies had more than 25 participants and only four studies included two or more datasets.
Conclusion: There is a need to assess multiple feature selection methods upon large cohort data consisting of multiple (3 or more) PA datasets. The cut-off criteria e.g. number of components, pairwise correlation value, explained variance ratio for PCA, etc. should be expressly stated along with any hyperparameters used in clustering.

1. Introduction

Accelerometer-based devices that measure raw accelerations are popular tools for physical activity researchers given their convenience and capacity to record data for several days to several weeks at a time. However, the data can be difficult to interpret in terms of meaningful outcomes. Identifying clusters from accelerometer data could facilitate quantification of types (modes) of specific physical behaviours associated with risk of serious health conditions, such as high levels of sedentary behaviour and/or low levels of purposeful physical activity, which might contribute to increased risk of chronic disease such as type 2 diabetes and cardiovascular disease [1]. This quantification would enable identification of the time spent and/or energy expended in target behaviours associated with risk of developing type 2 diabetes [2], other

* Corresponding author at: Origin Ward, Leicester Diabetes Centre, Leicester General Hospital, Gwendolen Road, Leicester LE5 4PW, UK.
E-mail address: pj100@leicester.ac.uk (P.J. Jones).

https://doi.org/10.1016/j.gaitpost.2021.08.007
Received 2 November 2020; Received in revised form 3 March 2021; Accepted 8 August 2021
Available online 13 August 2021

lifestyle diseases or conditions whose occurrence or progression is modulated by physical activity [e.g. 3], and premature all-cause and cardiovascular mortality [4].

1.1. Unsupervised vs supervised machine learning\

Machine learning concerns the use of mathematical algorithms either to find patterns in data (for example, habitual physical activity), divide data into clusters (for example by intensity of physical activity) or to identify data associated with a given output (for instance, accelerometer data associated with walking or sitting down). Machine learning algorithms are automated, learning from training data. Supervised machine learning, as the name implies, requires some human supervision – typically labelled data that it can learn from - for instance, all examples of walking in some accelerometer training data might be labelled with the same integer to enable the algorithm to use this label to learn the typical accelerometer data properties associated with it.

In contrast, unsupervised machine learning algorithms don't rely on human input and can work without the need for any labelled data or information on outcomes [5]. An unsupervised machine learning algorithm works on its own to uncover hidden patterns or clusters (grouping similar data together). Unsupervised machine learning techniques have the advantage of learning from accelerometer data to identify meaningful clusters representing different types/categories of physical activities. This avoids the need for provision of training data labelled with activity type, which can be both time-consuming and expensive to create.

1.2. Examples of unsupervised machine learning algorithms

K-means [6] is a popular clustering algorithm where the number of clusters is either known, presumed or indicated beforehand (choosing k is the main drawback of this algorithm although a number of techniques exist to derive possible values for k [7–9]). The centres of each cluster (centroids) are usually either initialised randomly or where the first centroid is initialized randomly then the other centroids are chosen in such a way as to be spread out as much as possible (known as K++, [10]). K-means then associates each data point x^i with the closest centroid μ_j. The set of points belonging to cluster i are shown as C_i:

$$C_i = \{x : x - \mu_i \leq \|x - \mu_j\|\}.$$

The algorithm's second step recalculates the centroids of each cluster to minimise the sum of squared Euclidean distances from the data points of this cluster to the cluster centroid.

$$\mu_j = \frac{1}{|C_j|}$$

where $|C_i|$ is the number of points in cluster. This two-step algorithm minimises the sum of squared Euclidean distances from each data point to the nearest centroid.

DBSCAN [11] is another example of an unsupervised algorithm which works by forming clusters around dense neighbourhoods. A point (X) with more than a minimum number of neighbours (N_{min}) within a given epsilon (ε) radius or neighbourhood (N) – i.e. the point's epsilon neighbourhood $N_\varepsilon(X)$ are known as core points (CP). The first such point is assigned to a new cluster.

$$N_\varepsilon(X) = \{j : X^j - X \leq \varepsilon\} CP = |N_\varepsilon(X)| \geq N_{\min}$$

Points which are density reachable (with a neighbour which itself is a neighbour of another core point) are also added to the same cluster. Core points which aren't density reachable are assigned to a new cluster and those points with sparse neighbourhoods ($<N_{\min}$) are treated as noise. As with the choice of k for K-means, there remains the difficulty of choosing optimal values for parameters like ε or $N_{\min}$ although various methods have been proposed [12,13].

The molecular complex detection algorithm (MCODE) [14] as its name suggests, was originally intended to cluster networks of interacting biomolecules and is density-based like DBSCAN. It uses a method of vertex weighting using a coefficient (c_i) useful for determining local neighbourhood density where ki is the vertex size of the neighbourhood of vertex i and n is the number of edges in that neighbourhood (i.e. the neighbourhood density of v not including v itself).

$$c_i = \frac{2n}{ki(ki-1)}$$

Vertices are weighted based on the highest k-core of the vertex neighbourhood where k-core is a graph of minimal degree K (graph g, for all v in g, $\deg(v) >= k$). The highest k-core of a graph is the central most densely connected subgraph. Using the vertex weighted graph as an input, the algorithm recursively moves outward from the seed vertex, including vertices whose weight is above a given threshold, given as a percentage away from the weight of the seed vertex (vertex weight percentage (VWP) parameter).

1.3. Importance of feature selection and extraction in physical activity accelerometer data clustering

An accelerometer feature is a numerical representation or function of the raw accelerometer values. There are hundreds of possible accelerometer features to choose from, for example, the dominant frequency from an accelerometer signal or its mean or maximum value for a given period of time. If there are not enough features, the model will perform poorly when trying to separate the data into clusters – for example, different physical activity intensities or types scattered evenly across clusters, so that we derive no utility from this sub-division of data. However, if too many features are used, clustering models can over-fit, limiting generalisability when applied to other datasets [15]. If a model has tailored itself too much to the data it learned from, this can negatively impact its performance on one or more new datasets (for instance if sedentary behaviour is now mixed up with moderate activity). A good model learns general widely applicable principles, which is why choice of feature is so important [16]. Too many features may also diminish the ability of the model to differentiate between physical activity clusters (the so-called 'curse of dimensionality' [17]). Therefore, a strategy for choosing features is necessary, firstly to reduce the complexity and computational burden of the clustering model – no-one wants to wait hours or days to process physical activity data, or not to be able to process it at all given insufficient processing power. Finding appropriate features can minimise the length of time required and the processing power needed by improving the efficiency of the machine learning model. It can also facilitate generalisation – applying the same model to other physical activity datasets and achieving similar clustering results. Finally, it enables us to better understand the role played by each feature and provides better model interpretability [18].

1.4. Examples of feature selection techniques

Feature selection methods aim to choose a small subset from the full set of original features found to be most useful in clustering physical activity. One example is correlation matrices which maps the relationship between features or between features and an output (for example cluster membership) [19]. The idea is that an appropriate feature subset contains features highly correlated with cluster membership yet relatively uncorrelated to each other.

Feature extraction also aims to reduce the number of features (dimensionality reduction) but does so by creating brand new features derived from the original feature set, for example by projecting those original features onto a new space with lower dimensionality [18] and then using these for clustering. An example of this is principal component analysis [20] which attempts to transform a large set of features into a smaller one whilst retaining as much of their information as possible. Each new feature is derived from a number of original features and the best new features (the principal components) are chosen.

The aim of this study is to perform a systematic review of studies which discuss accelerometer feature selection, for the purpose of assessing physical activity and sedentary behaviour. Our review includes peer-reviewed papers aiming to cluster accelerometer data into meaningful clusters of human physical activity type e.g. walking, running or by intensity (vigorous, moderate, light, sedentary) which employed a mathematical technique to either reduce the numbers of features selected or to evaluate subsets of features against each other. Additionally, we will catalogue the numbers of participants and datasets, and the type of accelerometer device used and in what wear location (e.g. wrist, hip, back, etc.).

2. Methods

2.1. Eligibility criteria

Studies were included in this review if they fulfilled the following selection criteria:

i. Theme: studies concerning unsupervised machine learning of human physical activity.
ii. Device: raw accelerations captured at multiple time points per second on at least three orthogonal axes using wearable accelerometer-based devices (regardless of accelerometer wear-site), or mobile phones.
iii. Types of machine learning: All unsupervised methods were included, examples of unsupervised methods include k-means clustering [6], or density-based clustering models such as DBSCAN [11].
iv. Physical activity: everyday physical activities and sedentary behaviours including studies limited to specific activities, e.g. walking, household activities, sedentary behaviour, cycling, running or taking the stairs.
v. Peer-reviewed papers published in English.
vi. Feature selection: employ a mathematical technique to either reduce the numbers of features selected or to evaluate subsets of features against each other.
vii. Clustering models based only on acceleration as opposed to combined models featuring gyroscopes, heartbeat and other sensors.

2.2. Information sources

Medline, Google Scholar, Scopus and Web of Science were searched papers published from any date until 21 January 2021. The search strings used to search Google Scholar and Scopus included the following:

ALL ((“unsupervised machine learning” OR “unsupervised clustering”) AND “physical activity” AND “accelerometer”)

ALL ((“machine learning” OR “clustering”) AND “physical activity” AND “accelerometer”) NOT “unsupervised”

The latter search was added to ensure articles that did not specifically mention ‘unsupervised’ were not missed.

The search string used to search Medline and Web of Science included the following:

“machine learning” AND “physical activity” AND “accelerometer”

The additional search of the Arxiv database was carried out on 8 February 2021 and utilised three search strings as follows:

“machine learning” AND “physical activity” AND “accelerometer”
“human” AND “movement” AND “accelerometer”
“clustering” AND “physical activity” AND “accelerometer”

2.3. Study selection

Data extracted include:

i. Article identification: title, authors and publication year.
ii. Methods: number and type of devices, sampling rate.
iii. Details/demographics: number of participants and physical activities, and participant demographics.
iv. Feature selection: method(s) employed to obtain features, list of features obtained.
v. Unsupervised clustering: machine learning method employed, and parameter information e.g. number of clusters (k), minimum number of points for a cluster ($N_{\min}$), etc.
vi. The number of datasets per study. Datasets were defined as the number of distinct cohort analyses in the study, datasets could differ by cohort (e.g. children and adults), context (e.g. lab versus free-living), device used to capture the data (e.g. brand of accelerometer).

The search data was stored in Excel and titles and abstracts screened for eligibility against the inclusion criteria by one reviewer (PJ). A second reviewer (AR) then independently assessed areas of ambiguity highlighted by the first reviewer by reviewing abstracts against the inclusion criteria. This systematic review methodology was registered retrospectively on the Open Science Framework on 23/02/2021 (see osf.io/8csjg/) (Fig. 1).

3. Results

3.1. Number of participant datasets and study sizes

The number of datasets per study varied between 1 dataset [22,25–31,33] and 5 datasets [23], with multiple datasets analysed in four studies [21,23,24,32] which consisted of 2 [21,32], 3 [24] and 5 datasets [23] (Table 2). Multiple datasets were employed to test features used for clustering physical activity across different cohorts of participant e.g. child, adult [23,24], within different environments e.g. laboratory or free-living [23], across different types of device [21] or to increase the quantity of participant data [22].

Total numbers of participants within studies varied from one person [29] to 160 adults [23] or 500 children [30]. Nine of the 13 studies had less than 25 participants [21,22,26–29,31–33], with the remaining four studies all above 131 participants [23–25,30]. Seven studies were undertaken in Europe [23,24,26–28,30,32], three in Asia [21,31,33], one in Australia [29], one in North America [25] and one unspecified [21].

3.2. Types of physical activity

The number of physical activity types within the datasets varied between 3 [33] and 33 [25] (Table 1). Six studies included seven activities or less [21,26–28,31–33], four included 9–12 activities [23,24,29,30], and two further studies involved 17 [22] and 33 [25] activities, respectively. The most common physical activities and sedentary behaviours within these studies included sitting ($n = 12$) [21–32], walking ($n = 11$) [21–28,31–35], running ($n = 8$) [22–24,27,28,31–33], standing ($n = 10$) [21–28,31,32] use of the stairs ($n = 8$) [21–28,31,32] (split into ascending and descending ($n = 6$) [21,22,25,26,31,32]), and lying ($n = 6$) [22–24,27–29]. The duration of time over which accelerometer data were collected varied from 1 h [29] to 2 days [30]. Three studies included free-living data [23,29,30] (See Supplementary Table 1).

3.3. Study demographics

Full details of age, gender and weight/height or Body Mass Index (BMI) were provided in only four of 13 studies [23–25,30]. Participants were aged between 20 and 30 in about half of these studies ($n = 6$) [21,22,26,31–33]. One study dataset consisted entirely of children aged 14 ($n = 500$) [30]. The remaining studies included ages 16–44 y [27,28], a mean age of 58.6 ± 17.4 y [25] or varied from 9 to 65 y [23,24] (Table 2). A gender breakdown was unavailable in eight of the 13 studies

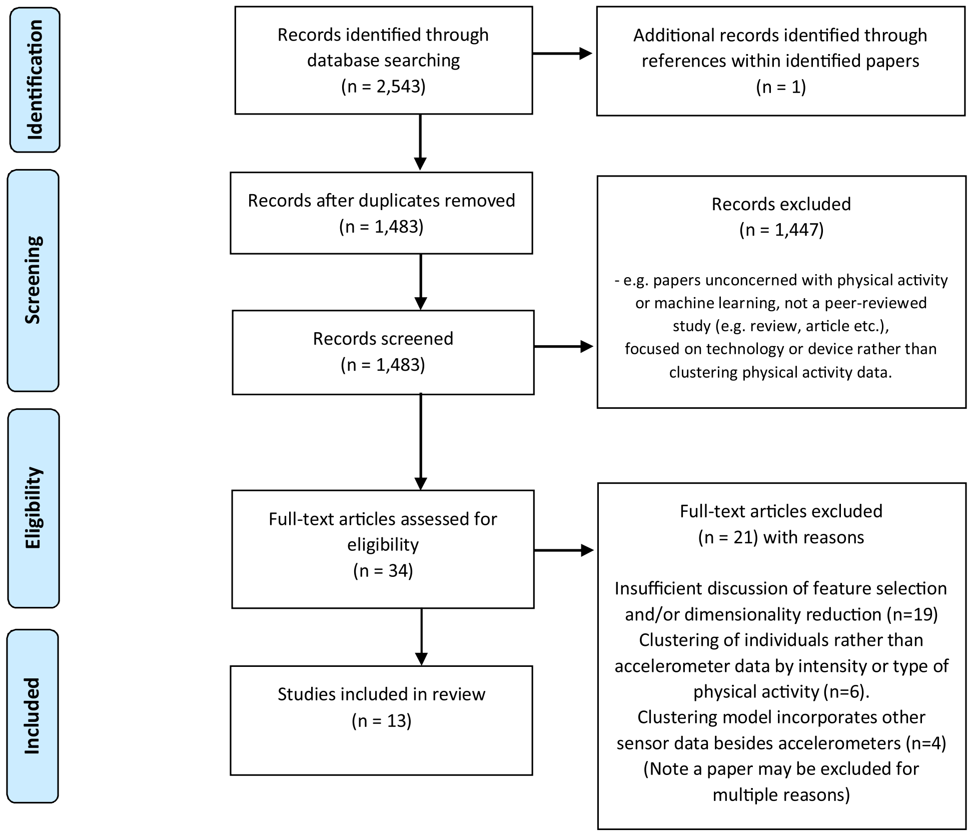

Fig. 1. PRISMA flow diagram of article selection process through the systematic review.

Table 1
Unsupervised clustering of physical activity studies.

Paper	Year	No activities	SLEEP	LYING	SIT	STAND	PERSONAL CARE	WALK	STAIR ASCEND	**STAIR DESCEND**	**RUN**	**Examples of other**
Dobbins [21]	2018	6, 6	–	–	x	x	–	x	x	x	–	Biking
He [22]	2018	17	–	x	x	x	–	x	x	x	x	Inclined treadmill running, cross-trainer, rowing.
Jones [23]	2019	9	–	x	x	x	–	x	x		x	Housework
Jones [24]	2020	9	–	x	x	x	–	x	x		x	Housework
Kheirkhahan [25]	2018	33	–	–	x	x	x	x	x	x	–	Housework, yoga, strength/weights, garden
Lago [26]	2019	6	–	–	x	x	–	x	x	x	–	Biking
Machado [27]	2014	7	–	x	x	x	–	x	–	–	x	Lying down (right/left side/belly up)
Machado [28]	2015	7	–	x	x	x	–	x	–	–	x	Lying down (right/left side/belly up)
Nguyen [29]	2007	12	x	x	x	–	x	–	–	–	–	Reading in bed, skipping, push-ups, driving
Van Kuppevelt [30]	2019	10	x		x	–	x	–	–	–	–	classwork, socialising, games, watching TV
Wang [31]	2016	6	–	–	x	x	–	x	x	x	x	–
Wang [32]	2017	6, 6	–	–	x	x	–	x	x	x	x	–
Yong [33]	2013	3	–	–	–	–	–	x	–	–	x	Throwing

Types of physical activity included within studies are marked by 'X' on the table above.
Van Kuppevelt includes a category of "sleeping and resting (including sick in bed)" [30].

[21,26–29,31–33].

3.4. Device types

Among the 13 physical activity studies included in the review, 9 studies included a research-grade accelerometer-based device (i.e. not smartphone or smartwatch) derived datasets [22–25,27–30,33], 4 studies analysed smartphone-derived physical activity data [21,26,31,32] (Table 3), and one study included data derived from both smartphone and smart watches [21]. The models of accelerometer-based device used to capture physical activity data varied greatly including GENEA and GENEActiv models ($n = 3$) [23,24,30], Motionplus ($n = 2$) [27,28] and a host of other devices – for example, YEI 3-Space sensor [33], MTx 3-DOF [22] or Actigraph GT3X [25]. The dynamic range of the GENEActiv is ± 8 g, the GENEA ± 6 g, the ActiGraph GT3X+ ± 6g, and the Alive Heart Monitor ± 2.7 g. The dynamic range was not specified for the other devices.

Placement of these accelerometers varied but typically included

Table 2
Demographics of study participants within unsupervised machine learning of physical activity data.

Paper	Year	Cohort size	Age	Gender (M/F)	Weight (kg)	Height (cm)	BMI (kg/m^2)	Country	Geographical area
Dobbins [21]	2018	9, 9	25–30	–	–	–	–	Korea	Asia
He [22]	2018	8	20–30	4/4	–	–	–	–	–
Jones [23]	2019	60 30 41 23 6/8	40–6520–409–1419–4220 –29	38/62% 27/73% 41/59% 30/70% 67/33% 37.5/62.5%	80.6 ± 11.6 69.2 ± 15.3 43.0 ± 11.2 73.7 ± 13.0 73.0 ± 17.1 65.4 ± 11.3	176.2 ± 6.2 169.4 ± 0.1 150.2 ± 13.3 172.7 ± 7.9 171.5 ± 10.9 166.8 ± 9.8	–	UK	Europe
Jones [24]	2020	60 30 41	40–6520–409 –14	38/62% 27/73% 41/59%	80.6 ± 11.6 69.2 ± 15.3 43.0 ± 11.2	176.2 ± 6.2 169.4 ± 0.1 150.2 ± 13.3	–	UK	Europe
Kheirkhahan [25]	2018	146	58.6 ± 17.4	49/97	74.8 ± 17.1	167.5 ± 9.2	25.9 ± 6.5	USA	North America
Lago [26]	2019	9	25–30	–	–	–	–	Denmark	Europe
Machado [27]	2014	8	16–44	–	–	–	–	Portugal	Europe
Machado [28]	2015	8	16–44	–	–	–	–	Portugal	Europe
Nguyen [29]	2007	1	–	–	–	–	–	Australia	Australia
Van Kuppevelt [30]	2019	500	14	42.4%/57.6%	–	–	21.3 ± 4.2	UK	Europe
Wang [31]	2016	6	22–28	–	–	–	–	China	Asia
Wang[a] [32]	2017	176	–22–28	—	—	—	—	Italy	Europe
Yong [33]	2013	5	20	–	–	–	–	Malaysia	Asia

[a] 17 of 30 subjects are selected from this UCI dataset.

Table 3
Data capture methodology for unsupervised machine learning of physical activity.

Paper	Year	Sets	Cohort size	Device(s)	No. devices	Model of device	Placement	Sampling rate (Hz)	Window size (s)	Window overlap
Dobbins [21]	2018	2	9 9	SmartphoneWatch	84	Samsung Galaxy S3 minis (2), S3s (2), and S+ (2) LG Nexus 4s (2) LG Watches (2) Samsung Galaxy Gears (2)	Waist (pouch) Wrist	50–200100–200	2	50%
He [22]	2018	1	8	Combined Accelerometer, Gyroscope, & Magnetometer	5	MTx 3-DOF	Torso, right arm, left arm, right leg, left leg	25	–	–
Jones [23]	2019	5	60 30 41 23 6/8	Accelerometer	2	GENEA GENEActiv GENEA GENEActiv GENEActiv	Left/Right Wrist	80–100	10	0%
Jones [24]	2020	3	60 30 41	Accelerometer	2	GENEA GENEActiv GENEA	Left/Right Wrist	80–100	10	0%
Kheirkhahan [25]	2018	1	146	Accelerometer	1	Actigraph GT3X	Right wrist	100	15	50%
Lago [26]	2019	1	9	Smartphone	8	Samsung Galaxy S3 minis (2), S3s (2), and S+ (2) LG Nexus 4s (2)	Waist (pouch)	50–200	2	50%
Machado [27]	2014	1	8	Accelerometer	1	Motionplus	Waist	800	4000 samples	–
Machado [28]	2015	1	8	Accelerometer	2	Motionplus	Wrist, waist	800	–	–
Nguyen [29]	2007	1	1	Heart monitor & Accelerometer	1	Alive Heart Monitor feat. Accelerometer	Waist	75	–	–
Van Kuppevelt [30]	2019	1	500	Accelerometer	1	GENEActiv	Wrist (Non-dominant)	40	–	–
Wang [31]	2016	1	6	Smartphone	1	Samsung Galaxy SIII	Front pants leg pocket	50	6 s	Overlap but unspec.
Wang [32]	2017	2	17 6	Smartphone	2	Unspecified	Unspecified	50	2.56 5.12	50%
Yong [33]	2013	1	5	Accelerometer	1	YEI 3-Space Sensor	Right arm (above)	–	–	–

Cohort size = number of participants participating in the study or each part of the study.

wrist ($n = 5$) [23–25,28,30] or waist ($n = 4$) [21,27–29], or alternatively the arm [33] or at numerous locations ($n = 1$) [22]. Smartphones ($n = 4$) included the Samsung Galaxy S3 ($n = 3$) [21,26,31], LG Nexus ($n = 2$) [21,26], with phone brand and model unspecified in one study [32]. Positions included waist [21,26], trouser/pants pocket [31], right arm [33] or were unspecified [32].

Sampling rates for the devices were usually within the 50–200 Hz interval ($n = 8$) [21,23–26,29,31,32], but sometimes as low as 25–40 Hz ($n = 4$) [22,30] or as high as 800 Hz [27,28] (Table 3). Among the research accelerometer studies ($n = 9$), sampling rates were 25–40 Hz [22,30], 75–100 Hz ($n = 4$) [23–25,29] or 800 Hz [22,23] and in one study was unspecified [33]. Smartphone sampling rates were within the region of 50–200 Hz [22,26,31,32]. The number of smartphone devices utilised to obtain data range from 1 [31] to 8 [21,26].

3.5. Methods of shortlisting features for evaluation

Many of these papers use the list of features in previously published studies to shortlist features for evaluation ($n = 7$) [21–25,29,32] or the decision-making process in shortlisting was not expressly addressed ($n = 5$) [26–28,30,33] (See Supp. Table 2). Sometimes features derived from a novel process were shortlisted. In one study, both time domain features from a prior study were shortlisted along with novel wavelet features intended to discriminate activities with similar amounts of energy by computing energy levels within a sub-band [22]. In another study, a 32-dimensional feature vector representing the rate of change of acceleration was proposed with the aim of enhancing the recognition rate of physical activity to ascend and descend stairs [31]. When previously published studies were used to derive features, there were only two studies that derived their shortlist from a mixture of supervised and unsupervised machine learning ($n = 2$) [23,24]. The remainder drew all their shortlisted features from supervised machine learning studies ($n = 3$) [21,22,32] or from a combination of supervised machine learning studies and studies unconnected with any form of machine learning ($n = 2$) [25,29].

3.6. Pre-processing applied to physical activity data

In some instances, filters were applied to physical activity data, for example a Butterworth filter was applied to the vector magnitude (vm) of raw accelerometer data ($vm = \sqrt{x^2 + y^2 + z^2}$) either to remove noise [21] or both noise and gravity from body acceleration [27,28]. Similarly, Butterworth filters were applied in two other studies [27,28] either with cut-off frequency of 0.25 Hz [27,28] or 3 Hz. [21]. Six studies expressly used normalisation, which took the form of min-max [23,24], division by maximum absolute value ($n = 1$) [21], division by mean ($n = 1$) [31], normalisation to zero mean and unit variance ($n = 2$) [27,28], or a normalisation term-frequency (inverse document frequency function) [25].

3.7. Shortlisted features

The total number of shortlisted features varied from 4 [30] to 54 [26] (Table 4). The number of shortlisted time domain features for evaluation ranged from 1 [30] to 10 [26] (Table 4). Typical shortlisted time domain features included the mean and standard deviation ($n = 9$) [21–24, 26–29,32], skewness and kurtosis ($n = 6$) [23,24,26–28,32] (Supplementary Table 3a). The number of shortlisted frequency domain features varied from 2 [23,24,29] to 8 [27,28] (Table 4). The most frequently shortlisted frequency domain features were the dominant frequency and power of dominant frequency ($n = 5$) [23–25,27,28], and zero crossing rate ($n = 4$) [22,27,28] (Supplementary Table 3b).

Correlation features appeared in 4 of 13 studies [22,27–29], and included pairwise correlation between axes e.g. XY, XZ, YZ [27,29] or autocorrelation (cross-correlation of a signal with a lagged version of itself over successive time intervals) [22]. Orientation angles of the acceleration axis relative to the horizontal plane derived from estimation of earth's gravity were also used as features in 5 studies [24–26,29,30] (Table 4) . Both formulae produce the same angle but one calculation is based on tangent and another on sine.

Van Kuppevelt et al. [30] used these angle features, while Jones et al. [23,24] extracted further features from them including minimum, maximum, mean, median and standard deviation of the orientation angle. Kheirkhahan et al. [25] also utilised mean and standard deviation of the orientation angle. In Nguyen, angle features were simply described as 'tilt angle' without further description [29]. Variations in calculation included Van Kuppevelt et al. [30] where for all continuous time periods with no z-angle change of more than 5 degrees lasting at least 5 min, the acceleration values were set to zero [30]. Finally, wavelet features were used in 1 study [22] (Table 4).

Table 4
Feature summary.

Paper	Year	Total No. of features	Time domain features	Frequency domain features	Correlation features	Angle features	Raw accelerometer data	Other feature types
Dobbins [21]	2018	8	5	3	0	–	–	
He [22]	2018	14	8	5	1	–	–	Wavelet
Jones [23]	2019	18	12	2	–	4	–	
Jones [24]	2020	18	12	2	–	4	–	
Kheirkhahan [25]	2018	7	2[a]	3	–	2	–	
Lago [26]	2019	54	10	–	–	–	–	Features extracted through codebooks, PCA, deep learning and matrix factorisation
Machado [27]	2014	18	8	8	2	–	–	
Machado [28]	2015	18	8	8	2	–	–	
Nguyen [29]	2007	10	3	2	1	1	3	
Van Kuppevelt [30]	2019	4	1[a]	–	–	3	–	
Wang [31]	2016	Second order derivative acceleration features						
Wang [32]	2017	11	8	3	–	–	–	
Yong [33]	2013	6	Original features are unspecified					

[a] Includes vector magnitude features.

3.8. Dimensionality reduction techniques

Among the studies which employ dimensionality reduction, the majority used one preferred technique ($n = 6$) [22–24,27,28,33], while others used 2 [21], 3 [25] or 4 methods [26]. These included a range of diverse techniques including PCA and variations of PCA ($n = 4$) [21,22,26,33], correlation feature selection (CFS) ($n = 3$) [21,23,24], and Best Cluster Permutation algorithm ($n = 2$) (producing a classification confusion matrix assessing each feature's contribution to maximising correct classification) [27,28]. Other approaches included three alternative methods of standard wrist model, bag of words representation, and supervised shape representation [25] (Table 5).

The number of principal components derived from the features ranges between 2 [21] and 100 [26] including 3 [33], 25, 50 and 100 [26] and in one instance was unspecified [22]. Correlation was also used to reduce the number of dimensions used to perform clustering e.g. to reduce features to a subset with zero correlation with each other [21], or to positively select the most relevant features [23,24].

3.9. Machine learning techniques

8 of 13 studies utilised the popular k-means algorithm [21–24,26–28,33] (Table 5). Other algorithms utilised for unsupervised include MCODE ($n = 1$) [31], hierarchical clustering ($n = 1$) [21] and DBSCAN ($n = 1$) [21]. Whilst k-means was a popular choice, its implementation varied. The number of clusters (k) was chosen in a variety of ways: either linked to the number of classes ($n = 3$) [23,24,29], derived from silhouette score ($n = 1$) [21], taking the first principal component to provide an agglomerative clustering of its coordinates before applying agglomerative clustering as an initial partition for k-means ($n = 1$) [22] or was unspecified ($n = 3$) [27,28,33]. Initialisation of k-means was through k++ ($n = 2$) [23,24], utilised agglomerative clustering for partitions ($n = 1$) [22] or was unspecified ($n = 6$) [21,26–28,33].

In one instance, k-means was used to cluster physical activity data using a metric other than Euclidean distance – dynamic time warping [26]. K-means was applied directly to the selected features, experimentally to pre-chosen subsets of features or used Principal Components [33] to derive a number of components.

In some instances, K-means was used in isolation to cluster physical activity by intensity [23,24], in others it formed part of a two-tier approach where a supervised classifier like the decision tree [33], random forest or support vector machine algorithms [26] were applied thereafter or it was used to cluster each individual subject's data rather than the data as a whole [28].

The DBSCAN algorithm is also used in one study [21]. This algorithm necessitates choosing an appropriate value for N_{min} (the minimum number of points necessary for a point to be considered 'core' and to belong to a cluster) and/or epsilon (the radius which defines the neighbourhood of a point in which a minimum number of points must exist for it to be considered core) but these parameters are not discussed further within this study [21].

3.10. Evaluating outcomes

Multiple methods (more than two) were used to evaluate the clusters produced using the selected features in 8 studies [21–26,30,31]. Popular methods included Adjusted Rand Index (ARI) ($n = 4$) [22,27,28,31], Average Cluster and Event Purity (ACEP) ($n = 3$) [23,24,29], Foulkes-Mallow Index (FMI) ($n = 2$) [22,31] and Silhouette Score ($n = 3$) [23–25] (Table 5). Sometimes evaluation methods were borrowed from supervised machine learning e.g. F1 score [25,26], average precision or recall [28] and confusion matrices [23,24,33] to show the contents within clusters.

The wide diversity in the features used to cluster accelerometer data when assessing physical activity, and the methods used to evaluate the features, currently make meta-analysis impossible and offers limited scope for directly comparing the effectiveness of different methods of dimensionality reduction to optimise unsupervised clustering of

Table 5
Feature selection for unsupervised machine learning in accelerometer physical activity research.

Paper	Feature selection method						
	Method 1	Method 2	Method 3	Method 1	Method 2	Other methods	Evaluation method
Dobbins [21]	PCA	Correlation	–	K-Means	DBSCAN	Hierarchical	Dunn Index, Distance Ratio, Entropy
He [22]	MPCA	–	–	Agglom + K-Means	–	ARI, RI, FMI	
Jones [23]	Correlation	–	–	K-Means	–	–	ACEP, Confusion Matrices, Silhouette Score
Jones [24]	Correlation	–	–	K-Means + FilterK	–	ACEP, Confusion Matrices, Silhouette Score	
Kheirkhahan [25]	Standard Wrist Model	Bag of Words	Sup. Shape Feature	Codebook	–	–	F1, Silhouette Score
Lago [26]	Evaluation of four sets of features (generated through codebooks, PCA deep learning & matrix factorisation)			K-Means + Dynamic Time Wrapping		–	F1, average precision and recall
Machado [27]	Best Cluster Permutation	–	–	K-Means	–	–	ARI
Machado [28]	Best Cluster Permutation	–	–	K-Means	Mean Shift	Affinity Propagation, Spectral Clustering	ARI
Nguyen [29]	Evaluation of four sets of features (raw vs. mixed, tilt/angle, signal mag.)			HMM-GMM	–	–	ACEP
Van Kuppevelt [30]	Evaluation of two sets of features (acceleration vs acceleration/angles)			HSMM	–	–	KLD, Correlation Matrices
Wang [31]	Evaluation of two sets of features (Novel vs. mixed domain)			MCODE	–	–	ARI, FMI
Wang [32]	Evaluation of three sets of features (Frequency vs. time vs. mixed domain)			Spectral	Single Link	Ward and Avg Link, K-Mediods	C-index
Yong [33]	PCA	–	–	K-Means	Fuzzy C-Means	–	Confusion Matrices

Key: ARI = Adjusted Rand Index, F1 = F1 Score, FMI = Foulkes-Mallow Index, JI = Jaccard Index, KLD = Kullback-Leibler Divergence, NMI = Normalized Mutual Information, NDR = no dimensionality reduction attempted.

physical activity.

4. Discussion

The aim of this systematic review was to summarise feature selection techniques applied in studies concerned with unsupervised machine learning of accelerometer-based device obtained physical activity, and to identify commonly used features identified through these techniques. Features were usually shortlisted on the basis of previous supervised machine learning studies rather than their actual performance in clustering. Typically, they included mean, standard deviation, skewness, kurtosis, dominant frequency and power of dominant frequency. PCA and correlation feature selection were the most commonly used techniques in feature selection. There was a wide variety of methods employed for evaluating clustering model performance.

There are a number of papers which have extracted features on the basis of previous studies or exploratory graphs of acceleration data, but were excluded as they lacked a formal means of evaluating alternative features or subsets of features, their relative importance or reducing them to those which are most important [34–37].

In some instances, techniques are suggested for manufacturing new features ("feature engineering") but there is little discussion of upscaling (compute power and processing time required to apply these techniques to larger datasets) or how they compare directly to traditional techniques (e.g. Fourier transforms to obtain frequency domain features) in terms of performance improvement versus runtimes. This is particularly important as the majority of these studies are based either upon small numbers of participants (less than 50) or single datasets, meaning their ease of application in clinically meaningful sample sizes is often unproven.

Engineered features may be more opaque as abstract 'black box' constructs which are less intuitive to visualise, along with their relationship to the original set of shortlisted input features. They may necessitate analysis of their characteristics and weightings or a horizon plot visualisation for each physical activity to gain insight [22,25,27,28].

Time and frequency domain features are usually shortlisted for use in unsupervised clustering of physical activity given their previous use in supervised classification problems. This systematic review provides a useful reference set of features used (Supp. Tables 3A and 3B) in unsupervised machine learning studies attempting to cluster physical activity which may be evaluated. Whilst the comparison of the relative performance of feature subsets provides useful data for unsupervised clustering of physical activity, there remains a need to identify the role of key specific features in clustering performance.

PCA remains the most popular choice for dimensionality reduction although it is interesting to note that PCA-derived cluster visualisations or explained variance ratio are usually not included. Few studies compare the efficiency of two or more methods of dimensionality reduction [21,25,26] in improving clustering and to date only one study has compared new methods against traditional ones such as PCA [26] albeit on the basis of F1, average precision and recall rather than internal measures evaluating cluster cohesion and separation. There is also a lack of consensus around the best methods to evaluate clustering. Assessment purely through classification techniques like average precision, recall, F1-measure etc. should be avoided because we cannot rely on the availability of labels particularly in free-living physical activity data and because internal measures such as silhouette score reveal the cohesion and separability of the clusters within the model used.

Limitations of this systematic review include the requirement that eligible articles be published in English which precludes relevant studies in other languages from being examined. We acknowledge that this review is limited by our search methodology that was restricted to broad terms ("physical activity", "accelerometer" and "unsupervised machine learning" or "unsupervised clustering") as well as our record selection criteria. There may be some studies which employ clustering algorithms but do not use some of these terms. Notably there were no studies identified using popular consumer-orientated accelerometer devices, e.g. iWatch, Fitbit. This reflects the lack of access to the raw acceleration data in such devices due to the 'black box' nature of the processing and generation of proprietary units. Populations studied were typically healthy, or the health status was unidentified, with limited data available on clinical and older populations. Finally, there is a dearth of data from studies undertaken in Africa or within low resource/indigenous communities where the acceleration signatures for physical activities may differ from high resource economies. The strengths of this study include the large number of papers brought back by these criteria (over 1,400 papers excluding duplicates) and the well-defined inclusion/exclusion criteria being applied to six databases.

Where unsupervised machine learning techniques are applied, it is very important to ensure all hyperparameters are recorded so that experiments are reproducible – for instance, the method employed to determine the number of clusters (k) for k-means, initialisation, and the distance metric applied. This is equally true for other algorithms such as DBSCAN which necessitate choices for epsilon or the minimum number of neighbouring points necessary for a point to be considered core etc. During our review, we found that these hyperparameters were often unreported. Where these cannot be included within a paper for reasons of brevity, the script should be made publicly available for ease of replication by other researchers looking to build upon earlier research.

To summarise, in datasets selecting features from accelerometer data to identify meaningful clusters that represent different types/categories of physical activities, we recommend that

- The range, scale and context of studies is extended. This is particularly important to facilitate the application of machine learning/deep learning in more diverse populations. Datasets should be annotated for the current health and functional status of the participants, accelerometer details (e.g. dynamic range, resolution, wear-site), and the context of the study, i.e. free-living, structured laboratory task.
- There is a need to assess multiple dimensionality reduction methods (both PCA, correlation, and newer techniques) upon a large cohort (100+ participants) consisting of multiple (3 or more) accelerometer datasets to establish a model performs consistently well.
- Cut-off criteria for dimensionality reduction e.g. number of components, pairwise correlation value, explained variance ratio for PCA etc. should be reported.
- Where new methods of feature creation are introduced, derived features should be subjected to comparative assessment alongside existing features e.g. time or frequency domain, angle orientation, etc.
- Feature assessment in labelled datasets should be through both internal and external measures of clustering quality – for example, silhouette score, Average Cluster and Event Purity and Confusion Matrices. The performance of feature sets chosen through alternative feature selection methods should also be assessed against required compute power and average run-times.
- Hyperparameter values for clustering algorithms should be published via appendices or open access scripts including methods for determining k for k-means, initialisation and the distance metric applied to ensure experiments can be reproduced.

5. Conclusions

Through this systematic review, we have identified a core set of features that studies to date have shortlisted for inclusion in unsupervised machine learning models for clustering physical activity data derived from raw acceleration accelerometers. The most popular methods for dimensionality reduction are PCA and correlation (either pairwise or autocorrelation). Cluster evaluation methods remain diverse, often consisting of methods used in classification which rely upon labels, while internal measures to evaluate the cohesiveness and separability of the clusters found are less common. Where these are used, Silhouette Score is most common although other measures such as Dunn Index are also used. The quality of these studies is extremely variable where some methodological details such as the method for choosing k in k-means or details of its initialisation and distance metric may be missing. Accordingly, we offer a suggested way forward for future research in this area looking to derive a shortlist of features for physical activity clustering and to compare alternate methods of feature selection.

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.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

The authors acknowledge the assistance of the University of Leicester Library Service for assistance in obtaining some of the papers necessary to carry out this review. We also thank the National Institute for Health Research (NIHR) Leicester Biomedical Research Centre, and the NIHR Collaboration for Leadership in Applied Health Research and Care–East Midlands for their consent and encouragement for this review.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.gaitpost.2021.08.007.

References

[1] B.J. Jefferis, P.H. Whincup, L. Lennon, S.G. Wannamethee, Longitudinal associations between changes in physical activity and onset of type 2 diabetes in older British men: the influence of adiposity, Diabetes Care 35 (2012) 1876–1883, https://doi.org/10.2337/dc11-2280.
[2] G. Hu, Q. Qiao, K. Silvetoinen, J.G. Eriksson, P. Jousilahti, J. Lindström, et al., Occupational, commuting, and leisure-time physical activity in relation to risk for type 2 diabetes in middle-aged Finnish men and women, Diabetologia 46 (2003) 322–329, https://doi.org/10.1007/s00125-003-1031-x.
[3] K. Okkersen, C. Jimenez-Moreno, S. Wenninger, F. Daidj, J. Glennon, S. Cumming, et al., Cognitive behavioural therapy with optional graded exercise therapy in patients with severe fatigue with myotonic dystrophy type 1: a multicentre, single-blind, randomised trial, Lancet Neurol. 17 (8) (2018) 671–680, https://doi.org/10.1016/S1474-4422(18)30203-5.
[4] U. Ekelund, J. Tarp, J. Steene-Johannessen, B.H. Hansen, B. Jefferis, M. W. Fagerland, et al., Dose-response associations between accelerometry measured physical activity and sedentary time and all-cause mortality: Systematic review and harmonised meta-analysis, Br. Med. J. 366 (2019) I4570, https://doi.org/10.1136/bmj.l4570.
[5] S. Raschka, Python Machine Learning, Packt, Birmingham, UK, 2016 pp.1–4, 7–12 (Chapter 1).
[6] S.P. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (1982) 129–137.
[7] R.L. Thorndike, Who belongs in the family? Psychometrika 18 (4) (1953) 267–276.
[8] P. Rousseeuw, A graphical aid to the interpretation and validation of cluster analysis, Comput. Appl. Math. 20 (1987) 53–65.
[9] T. Calinsky, J. Harabasz, A dendrite method for cluster analysis, Commun. Stat. (1972) 1–27.
[10] D. Arthur, S. Vassilvitskii, K-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, Soc. Ind. Appl. Math. (2007) 1027–1035.
[11] M. Ester, H. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: E. Simoudis, J. Han, U.M. Fayyad (Eds.), Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, 1996, pp. 226–231.
[12] A. Sawant, Adaptive methods for determining DBSCAN parameters, IJISET 1 (4) (2014) 329–335.
[13] J. Esmaelnejad, J. Habibi, S.H. Yeganeh, A novel method to find appropriate ε for DBSCAN, ACIIDS: Part 1. LNAI 5990 (2010) 93–102.
[14] G.D. Bader, C.W.V. Hogue, An automated method for finding molecular complexes in large protein interaction networks, BMC Bioinform. 4 (1) (2003) 2.
[15] P. Domingos, A few useful things to know about machine learning, Commun. ACM 55 (10) (2012) 78–87, https://doi.org/10.1145/2347736.2347755.
[16] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271.
[17] R.E. Bellman, Dynamic Programming, Princeton University Press, Princeton, 1957.
[18] C.C. Aggarwal, C.K. Reddy, Data Clustering – Algorithms and Applications, CRC Press, Minneapolis, 2014. Paragraph 2.1.2, pp. 29-32.
[19] R. Vishal, Feature selection – correlation and p-value. Towards Data Science, 2018 https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf.
[20] J. Shlens, A tutorial on principal components analysis, 2005. Available at: https://www.cs.cmu.edu/~elaw/papers/pca.pdf (accessed 17.02.21).
[21] C. Dobbins, R. Rawassizadeh, Towards clustering of mobile and smartwatch accelerometer data for physical activity recognition, Informatics 5 (2018) 29, https://doi.org/10.3390/informatics5020029.
[22] H. He, Y. Tan, W. Zhang, A wavelet tensor fuzzy clustering scheme for multi-sensor human activity recognition, Eng. Appl. Artif. Intell. 70 (2018) 109–122, https://doi.org/10.1016/j.engappai.2018.01.004.
[23] P. Jones, E.M. Mirkes, T. Yates, C.L. Edwardson, M. Catt, M.J. Davies, K. Khunti, et al., Towards a portable model to discriminate activity clusters from accelerometer data, Sensors 19 (2019) 4505, https://doi.org/10.3390/s19204504.
[24] P. Jones, M.K. James, M.J. Davies, K. Khunti, M. Catt, T. Yates, et al., FilterK: A new outlier detection method for k-means clustering of physical activity, J. Biomed. Inform. 104 (2020) 103397, https://doi.org/10.1016/j.jbi.2020.103397.
[25] M. Kheirkhahan, A. Chakraborty, A.A. Wanigatunga, D.B. Corbett, T.M. Manini, S. Ranka, Wrist accelerometer shape feature derivation methods for assessing activities of daily living, BMC Med. Inform. Decis. Making 18 (Suppl. 4) (2018) 124, https://doi.org/10.1186/s12911-018-0671-1.
[26] P. Lago, S. Inoue, Comparing feature learning methods for human activity recognition: Performance study in new user scenario, Joint 8th International Conference on Informatics, Electronics and Vision (ICIEV) (2019) https://ieeexplore.ieee.org/abstract/document/8858548 (accessed 23.09.20).
[27] I. Machado, R. Gomes, H. Gamboa, V. Paixao, Human activity recognition from triaxial accelerometer data – feature extraction and selection methods for clustering of physical activities, Proceedings of the International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS-2014) (2014) 155–162, https://doi.org/10.5220/0004749801550162.
[28] I.P. Machado, A.L. Gomes, H. Gamboa, V. Paixao, R.M. Costa, Human activity data discovery from triaxial accelerometer sensor: non-supervised learning sensitivity to feature extraction parametrization, Inf. Process. Manage. 51 (2015) 204–214, https://doi.org/10.1016/j.ipm.2014.07.008.
[29] A. Nguyen, D. Moore, I. McCowan, Unsupervised clustering of free-living human activities using ambulatory accelerometry, in: Proceedings of the 29th Annual international Conference of the IEEE EMBS Cité Internationale, Lyon, France, August 23–26, 2007.
[30] D. Van Kuppevelt, J. Heywood, M. Hamer, S. Sabia, E. Fitzsimons, V. van Hees, Segmenting accelerometer data from daily life with unsupervised machine learning, PLoS One 14 (2019) e0208692, https://doi.org/10.1371/journal.pone.0208692.
[31] D. Wang, L. Liu, X. Wang, Y. Lu, A novel feature extraction method on activity recognition using smartphone, WAIM 2016 Workshops, LNCS 9998 (2016) 67–76.
[32] X. Wang, Y. Lu, D. Wang, L. Liu, H. Zhou, Using Jaccard distance measure for unsupervised activity recognition with smartphone accelerometers, APWeb-WAIM 2017 Workshops, LNCS 10612 (2017) 74–83, https://doi.org/10.1007/978-3-319-69781-9_8.
[33] C.Y. Yong, N.H. Mahmood, R. Sudirman, K.M. Chew, Motion classification using proposed principle component analysis hybrid k-means clustering, Engineering 5 (2013) 25–30, https://doi.org/10.4236/eng.2013.55B006.
[34] N. Alshurafa, B. Mortazavi, W. Xu, Designing a robust activity recognition framework for health and exergaming using wearable sensors, IEEE J. Biomed. Health Inform. 18 (2013) 1636–1646, https://doi.org/10.1109/jbhi.2013.2287504.
[35] Y. Kwon, K. Kang, C. Bae, Unsupervised learning for human activity recognition using smartphone sensors, Expert Syst. Appl. 41 (2014) 6067–6074, https://doi.org/10.1016/j.eswa.2014.04.037.
[36] Y. Lu, Y. Wei, L. Liu, J. Zhong, L. Sun, Y. Liu, Towards unsupervised physical activity recognition using smartphone accelerometers, Multimed. Tools Appl. 76 (2017) 10701–10719, https://doi.org/10.1007/s11042-015-3188-y.
[37] Y. Wei, L. Liu, J. Zhong, Y. Lu, L. Sun, Unsupervised race walking recognition using smartphone accelerometers, in: S. Zhang, M. Wirsing, Z. Zhang (Eds.), Knowledge Science, Engineering and Management, Springer, 2015, pp. 691–702.