Data Mining and Knowledge Discovery (2022) 36:477–512
https://doi.org/10.1007/s10618-021-00817-w

A recurrent neural network architecture to model physical activity energy expenditure in older people

Stylianos Paraschiakos[1,2] · Cláudio Rebelo de Sá[3] · Jeremiah Okai[2] · P. Eline Slagboom[1] · Marian Beekman[1] · Arno Knobbe[2]

Received: 20 May 2020 / Accepted: 29 November 2021 / Published online: 10 January 2022

Abstract

Through the quantification of physical activity energy expenditure (PAEE), health care monitoring has the potential to stimulate vital and healthy ageing, inducing behavioural changes in older people and linking these to personal health gains. To be able to measure PAEE in a health care perspective, methods from wearable accelerometers have been developed, however, mainly targeted towards younger people. Since elderly subjects differ in energy requirements and range of physical activities, the current models may not be suitable for estimating PAEE among the elderly. Furthermore, currently available methods seem to be either simple but non-generalizable or require elaborate (manual) feature construction steps. Because past activities influence present PAEE, we propose a modeling approach known for its ability to model sequential data, the recurrent neural network (RNN). To train the RNN for an elderly population, we used the growing old together validation (GOTOV) dataset with 34 healthy participants of 60 years and older (mean 65 years old), performing 16 different activities. We used accelerometers placed on wrist and ankle, and measurements of energy counts by means of indirect calorimetry. After optimization, we propose an architecture consisting of an RNN with 3 GRU layers and a feedforward network combining both accelerometer and participant-level data. Our efforts included switching mean to standard deviation for down-sampling the input data and combining temporal and static data (person-specific details such as age, weight, BMI). The resulting architecture produces accurate PAEE estimations while decreasing training input and time by a factor of 10. Subsequently, compared to the state-of-the-art, it is capable to integrate longer

Responsible editor: Panagiotis Papapetrou.

✉ Stylianos Paraschiakos
s.paraschiakos@lumc.nl

1 Molecular Epidemiology, Department Biomedical Data Science, LUMC, Leiden, The Netherlands

2 Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands

3 Data Science research group, University of Twente, Enschede, The Netherlands

activity data which lead to more accurate estimations of low intensity activities EE. It can thus be employed to investigate associations of PAEE with vitality parameters of older people related to metabolic and cognitive health and mental well-being.

Keywords Recurrent neural networks · Physical activity energy expenditure · Accelerometer · Wearables · Indirect calorimetry · Monitoring older adults

1 Introduction

At older age, the extension of health span and maintenance of mobility are of great importance for the quality of life. Regular physical activity (PA) of moderate intensity is known to offer positive effects on the reduction of disease incidence and mortality risk (Manini et al. 2006; Chen et al. 2012; Cicero et al. 2012; Petersen et al. 2012). To quantify and monitor the intensity of PA, estimation of energy expenditure during physical activity is an obvious necessity. By monitoring physical activity energy expenditure (PAEE), older people may better engage in physical activities, leading to better health and reduced (multi)morbidity and mortality risk (Manini et al. 2006).

PAEE is one component of total energy expenditure (TEE), where TEE is the sum of PAEE, resting energy expenditure (REE or RMR) by a fasted individual, and thermic effect of food (TEF). One way to measure PAEE is using direct calorimetry and measurements of heat production, but expensive equipment is required. Also, the Doubly Labeled Water Technique (DLW) provides an accurate technique of TEE estimation from where PAEE can be estimated, however, similar to direct calorimetry, it requires sophisticated lab-based equipment to analyse urine samples. Therefore, indirect calorimetry (Leonard 2012) is commonly used, which involves the measurement of oxygen and carbon dioxide exchange by ventilated mask or hood.

Because such forms of calorimetry cannot be performed under free-living conditions, methods to estimate PAEE from wearable accelerometers have been developed (Lyden et al. 2011; Staudenmayer et al. 2009; Ellis et al. 2014; Montoye et al. 2017a; Caron et al. 2020; O'Driscoll et al. 2020). This form of indirect calorimetry is estimated by accelerometer data and their combinations with physiological measurements such as heart rate, and individual-level data (demographic, anthropometric) using both linear and non-linear methods (Liu et al. 2012). For example, linear or multiple regression methods can be used to estimate PAEE (Lyden et al. 2011), but also non-linear ensembles like random forest regressors (Gjoreski et al. 2013; Ellis et al. 2014; O'Driscoll et al. 2020) and deep learning method such as artificial neural networks (ANN) (Staudenmayer et al. 2009; Montoye et al. 2017a) and convolutional neural networks (CNN) (Zhu et al. 2015) have been employed. Good estimates of PAEE can be derived from accelerometry data.

Since the majority of currently available methods to estimated PAEE from accelerometer data are mainly developed and tested on a young or middle-aged population (Montoye et al. 2017a; Caron et al. 2020), these models may not be suitable for estimating PAEE among the elderly. This is due to the fact that the elderly differ in energy requirements (Roberts and Dallal 2005; Hortobágyi et al. 2003), expenditure (Frisard et al. 2007; Knaggs et al. 2011), and range of physical activities (Jones et al.

2009; Martin et al. 2014) while it is also seen that the older the individuals tent to spend more time in sedentary activities (van Ballegooijen et al. 2019).

There are two main drawbacks in the currently available methods. First, while linear models are pretty simple to deploy and use, they are unable to fit to all the activities (van Hees et al. 2009). Second, the non-linear can be quite elaborate and computationally intensive, since they require steps of features construction and selection in order to capture the temporal nature of the accelerometer signal. Thus, a PAEE modeling method that does not require any sophisticated or handcrafted pre-processing is called for, in addition to the development and testing of the model on older adults.

Therefore, we propose a neural network modeling approach that is known for its ability to model sequential data, the Recurrent Neural Network (RNN). The RNN is a network architecture that can deal with raw sensor data or minimum feature extraction, and can model temporal data by sequential processing. The nature of the processing in RNNs provides the possibility to remember information from the near as well as distant past, which is an advantage in comparison to ANN or CNN. Because past activities influence present PAEE, RNN modeling seems to be an excellent fit.

To train the RNN for application on an elderly population, we used the Growing Old Together Validation (GOTOV) dataset (Paraschiakos et al. 2020) with 34 healthy participants of 60 years and older (mean 65 years old), performing 16 different physical activities. This dataset is one of the first datasets publicly available with a focus on physical activity modeling of the elderly, both for activity recognition and energy expenditure. It includes multiple sensors (accelerometry, indirect calorimetry, physiological measurements) placed at multiple body locations. In the current study, we used a combination of accelerometers placed on wrist and ankle (GENEActiv), because accelerometers combined on hand and foot can be good PAEE estimators (Dong et al. 2013; Ellis et al. 2014). Furthermore, Montoye (Montoye et al. 2017a, b) argues that both wrist and ankle separately produce the best PAEE estimations. Finally, the measurements of energy counts (per-breath calories) were collected by means of the medical-grade COSMED device (McLaughlin et al. 2001).

Our proposed RNN architecture exploits Gated Recurrent Units (GRU) layers combined with a shallow ANN in order to make use of both accelerometer and participant-level data (age, gender, weight, height, BMI). This means that both temporal data and attribute-value data are given as input to the model and it combines them to give estimates of PAEE. In more detail, the model takes as an input sequences of temporal data representing a time window of past accelerometer, and creates output-features that are combined with the participant-level data in order to produce a PAEE estimation.

Summarising, the main contribution of this paper is the development of a novel PAEE modeling architecture without any sophisticated feature construction step focused on a population group that is often overlooked: adults over 60 years of age. The specific contributions of our work are the following:

1. We propose an original GRU-based approach for modeling PAEE, and demonstrate its efficacy in an elderly population. Once before, an RNN-based approach has been used for PAEE estimation (Mardini et al. 2020) combining LSTM and CNN layers.

While we will demonstrate that LSTMs work equally well on our dataset, we have adopted GRUs for reasons of higher efficiency.
2. We prove that using statistical dispersion metrics like standard deviation to down-sample the accelerometer data can significantly improve the accuracy achieved, while reducing the training time by approximately 10 times while using 10 times less data, compared to averaging (mean).
3. We show that longer windows of prior sensor (up to 2 min) lead to better PAEE estimation, and that GRU model based on standard deviation can deal with these longer windows efficiently.
4. We demonstrate how the addition of participant-level data (for example age and weight of a subject) can improve the sensor-based model.

The rest of the paper is structured as following. Section 2 presents the related work, while Sect. 3 presents the dataset used for model development. Then, Sect. 4 discusses the methodological steps needed to model PAEE, such as model architecture, data preparation (including the predictors down-sampling steps), model evaluation and experimental pipeline. This is followed by the results section (Sect. 5) presenting the main findings of our analysis. Finally, our findings, modeling strengths and limitations, and our future work is discussed in Sect. 6.

2 Related work

In the past few years, multiple PAEE methods have been developed, ranging from simple linear regression and linear mixed models (Montoye et al. 2017a) to non-linear ones, based on machine learning (Montoye et al. 2017a; Ellis et al. 2014; Zhu et al. 2015). Here, we give a short introduction of these by examining their modeling aspects in detail. Table 1 displays the three publications explained in this sections and their modeling set-ups.

Montoye et al. (2017a) already provided an interesting comparison of multiple PAEE methods. In this work, a linear regression model (LM) was compared to a linear mixed model (LMM), and a shallow artificial neural network (ANN). These models were developed in a dataset of $N = 40$ healthy participants ($\approx 50\%$ female) between ages 18 and 44 years (mean = 23.7). The dataset included recordings from 4 different accelerometers on the right hip, right thigh and both wrists, while a portable metabolic analyser on their backs connected to a breathing mask. The participants performed a 90-min semi-structured protocol of 13 activities of different intensity levels, such as lying down, sitting, household, climbing stairs, walking, jogging, stationary cycling and others in order and duration as determined by the participant.

In order to train the different models, time-domain predictor features of 30 non-overlapping seconds were developed per device. The features were chosen based on previous work of the authors and included: mean, standard deviation, minimum, maximum, co-variance of adjustment windows, and 10th, 25th, 50th, 75th and 90th percentiles per acceleration axis (triaxial accelerometers were used). While, as a target variable, they used 30 s of aggregated MET[1] values, synchronous with the predictors.

[1] Metabolic Equivalent of Task, where 1 MET at rest (e.g. sitting) equals 1 Kcal/kg/h.

Table 1 State-of-the-art methods and their characteristics

Publication	Montoye et al. (2017a)	Ellis et al. (2014)	Zhu et al. (2015)
N	40	40	30
Age	23.7	35.8	27.8
Body Location	Hip, thigh, wrists	Hips, wrist	Waist
Features	36 time domain	45 time and freq-domain	–
Window	30 s	1 min	5.12 s
Anthropometrics	False	False	True
Methods	LM, LMM, ANN	RF	CNN

Based on these, the different models (LM, LMM, ANN) were trained per device, where the two different ANNs developed were based on prior work of Staudenmayer et al. (2009).

The models were compared per body location using Pearson correlations, root mean squared errors (RMSE) and bias. The model correlations ranged from 0.82 to 0.89 and RMSE ranged from 1.07 to 1.31 MET for the four accelerometers with the ANN models, whereas the linear models (LM and LMM) from 0.71 to 0.88 and RMSE ranged from 1.11 to 1.61 MET. The differences between the ANN and the linear models (LM and LMM) were statistically significant for the wrists while for the thigh there was no significant difference for all models and for the hip only one of the ANNs had higher correlations and lower RMSE than the linear models.

Similar to Montoye et al. (2017a) and Ellis et al. (2014) developed a random forest regressor (RFr) and compared it to the ANN approach of Staudenmayer et al. (2009). This time, a broader set of features is used, N= 45, with both time- and frequency-domain features computed from non-overlapping windows of 1 minute. The majority of these features were aggregations of signal vector magnitude (SVM $= \sqrt{x^2 + y^2 + z^2}$) and angular features that capture orientation information of the accelerometer. Adding to that, participants wore a heart rate (HR) monitor and an extra feature of HR is included. The dataset included recordings from $N = 40$ healthy participants ($\approx$ 50% female) with mean age $=$ 35.8 years, wearing two accelerometers (both hips and dominant wrist) and a portable indirect calorimeter. Participants performed 3 household activities (out of a set of 5) and 3 locomotion activities (slow walk, brisk walk, treadmill jog) for 6 min each.

The RFr model was developed by learning 500 regression trees with a minimum leaf size of 5 using MET values as a target. Then, the predictions were evaluated on the minute level using the leave-one-subject-out (LOSO) cross validation procedure by measuring the bias, standard error and the RMSE. During the experiments, Ellis et al. compared how the models perform for a single body location, but also by combining them, and by adding HR data. In terms of RMSE, the RFr approach outperforms the ANN ones with an RMSE= 1.00 versus an RMSE between 1.12 and 1.35. Furthermore, about the body-locations, placing an accelerometer on the right or left hip produces similar performance to the wrist. However, when wrist and right hip are combined, the performance improves significantly compared to the single body

location. Finally, when HR data are included, both for one the body location set-up or for multiple, the performance further improved significantly.

The above work exploits handcrafted features in order to estimate PAEE, while the recent advances in deep learning give us the advantage to automate this procedure and extract complex features of sensor data while training. Zhu et al. (2015), such a method is introduced where a convolution neural network (CNN) architecture is used on a dataset of N = 30 healthy subjects ($\approx$ 33% female) with ages between 19 and 45 years (mean age = 27.8). The subjects performed a 30-min protocol of 6 activities (walking, climbing stairs, running, static standing/sitting, riding elevator) inside and outside a regular hospital facility. During the data collection, the subjects were equipped with a smartphone with triaxial accelerometer, placed in a waist pouch, and a portable indirect calorimeter that also records HR. Additionally, anthropometric features (height, weight, age, gender, etc.) per participant were included in the modeling procedure.

Before training, the triaxial accelerometer data (50 Hz) were transformed into sequences of 256 samples representing a time window of 5.12 s. As target data, the output of the indirect calorimeter (Kcal/min) was used, aggregated to the same rate as the accelerometer sequences. The trained CNN consisted of 2 convolution layers connected with one dense layer that takes as input the concatenation of the CNN features with the anthropometric data. The first CNN layer employs 8 filters of kernel size 5 and a pooling factor of 2, while the second CNN layer has 4 filters of size 5 and the same pooling factor, and the dense layer had a size of 400 nodes. As activation functions, both CNN layers used *tanh* while for the dense layer no activation function is used (linear transformation layer). Unfortunately, other important hyperparameters like the number of epochs and batch sizes during training, were not reported.

Their model was evaluated with LOSO cross-validation by measuring the RMSE and it was also compared to an activity-specific linear regression model and an ANN approach using handcrafted features. Overall, their CNN approach shows the lowest RMSE ($\approx$ 1.12) while the activity-specific one follows with RMSE = 1.59 and the ANN with RMSE = 1.79. When the models are tested per clusters of activities still, the CNN clearly outperforms both models in every activity cluster.

Concluding, different statistical or machine learning seem to estimate PAEE quite well. However, it is quite challenging to compare their reported performances since they are developed on (1) different datasets, (2) using data from accelerometers on different body locations (hip, thigh, waist, wrist), and (3) down-sampled these to different windows (from 5 s to 1 min). For this reason, in Sect. 5.5, we try to fairly compare all the above including our proposed method using a similar settings. In order to cope with the aforementioned challenges, we will develop all the methods using the same dataset [GOTOV (Paraschiakos et al. 2020)] with accelerometers on the ankle and wrist.

3 Dataset

The dataset used for our experiment is part of the Growing Old Together Validation (GOTOV) study. The GOTOV dataset is designed to develop both activity recognition

(Paraschiakos et al. 2020; Okai et al. 2019) and energy expenditure models that will serve multiple free-living ageing studies with similar population and devices (van de Rest et al. 2016; Westendorp et al. 2009; Wijsman et al. 2013). The dataset includes calorimetry measurements combined with the ankle and wrist accelerometer, among other data and since June 2020, is freely available in the 4TU data repository.[2]

3.1 Study population

The participants in the GOTOV study responded to advertisements on bulletin boards in public spaces in the city of Leiden, the Netherlands. People were eligible to participate in the study if they:

1. were older than 60 years old.
2. had a healthy to overweight BMI[3] between 23 and 35 kg/m^2.
3. had no restrictions in their movement caused by health conditions.
4. owned and had access to their own bicycle.

A total of 35 individuals (14 female, 21 male) between the ages 60 and 85 years old (mean 65) and mean BMI 27 kg/m^2 were recruited. Besides age, gender, height and weight, no additional clinical information was recorded on the participants. The GOTOV study was approved by the Medical Ethical Committee of LUMC (CCMO reference NL38332.058.11).

3.2 Data collection protocol

The 35 participants performed a set of 16 activities according to a specific protocol of approximately 90 min. The 16 activities were performed successively for specific time windows and with short breaks of standing still in between (1 minute). A researcher monitored the activities duration without giving any instructions or illustrations of the activities and wrote down their starting and ending timestamp. The activity protocol took place at two locations; *indoors* and *outdoors* of the Leiden University Medical Center (LUMC) facilities. The indoor activities consisted of *lying down, sitting, standing, walking stairs* and several household activities, such as *dish washing, staking shelves* and *vacuum cleaning*. The indoor activities were performed in a room equipped with all the necessary instrumentation. The outdoor activities included different types of walking *slow, normal, fast*, as well as *cycling*. A visual example of the procedure can be found in a recorded video.[4] The detailed protocol of the activities performed is described in Table 2, have in mind that between every two activities there was a break of 60 s standing, but in Table 2 this is represented only once, at the second row. Other than that, due to adverse weather conditions, only 25 out of 35 participants were able to perform the outdoor activities (walking, cycling).

[2] DOI link: https://doi.org/10.4121/12716081.

[3] Body-Mass Index, the body mass divided by the square of the body height.

[4] https://youtu.be/jvx5FGhqPxw.

Table 2 Activities and their duration performed in the GOTOV protocol

Activity (s)	Description
Light jumping (20 s)	Synchronising sensors
Standing (60 s)	Get some rest between activities
Steping (60 s)	A step test on a step (∼ 20 times)
Sedentary activities (180 s each)	1. Lying down (left and right)
	2. Sitting on a sofa (legs on)
	3. Sitting on a sofa (legs on the ground)
	4. Sitting on a chair while working
Walking stairs up (20 s)	Ascending two flights of stairs
Household activities (180 s each)	1. Washing dishes (while standing)
	2. Stacking shelves with books
	3. Vacuum cleaning
Walking outside (300 s each)	1. Slow pace
	2. Medium pace
	3. Fast pace
Cycling outside (900 s)	Participant's pace in normal traffic conditions

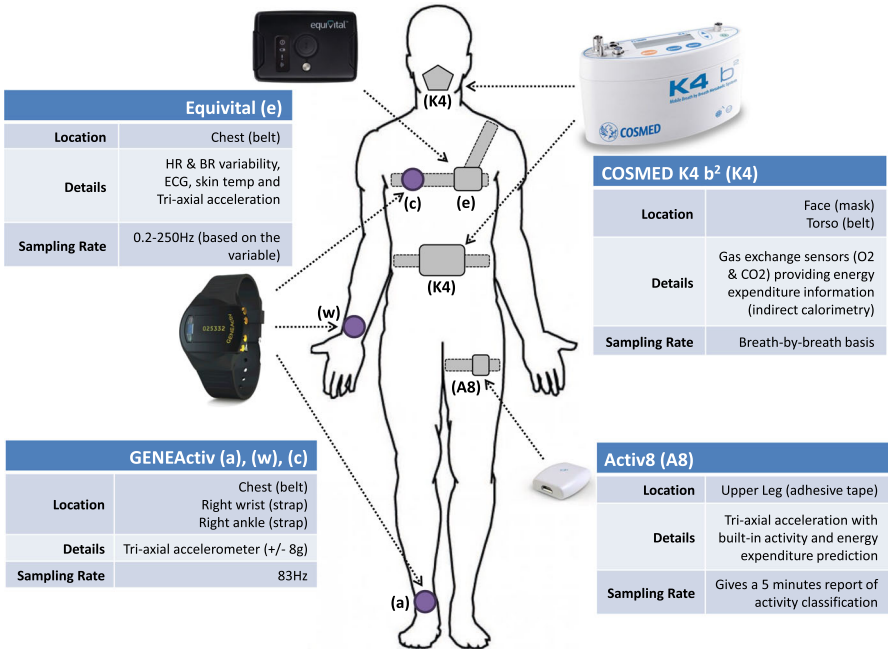

Fig. 1 GOTOV study devices and their body location (Paraschiakos et al. 2020)

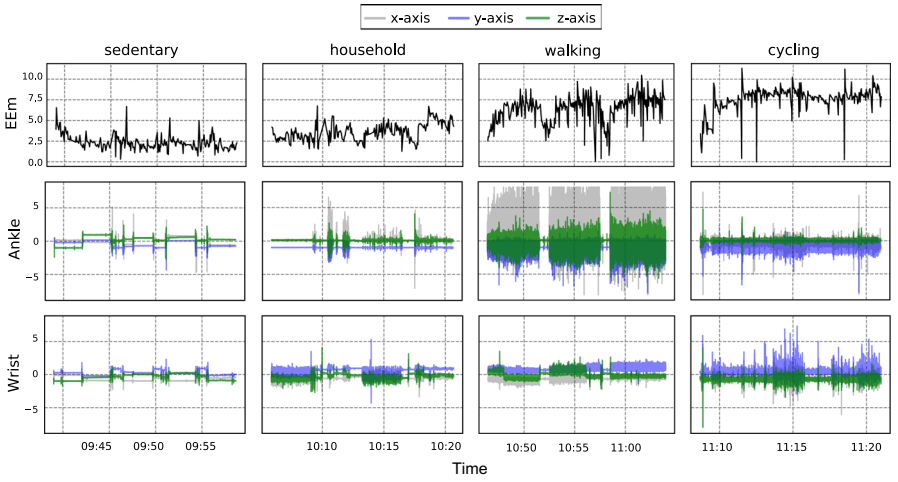

Fig. 2 Model input per device for the different activity groups

3.3 Devices and body locations

During the data collection, the participant used 4 different devices in 6 body locations (see Fig. 1). The set of devices included both accelerometers and sensors measuring physiological indicators, e.g. indirect calorimetry (VO_2, VCO_2), breathing rate (BR) and heart rate (HR). In this study, we focus on the data coming from accelerometers and indirect calorimetry. This is mainly motivated by the fact that the models will serve existing free-living studies using the same sensor setup.

Accelerometry The GENEActiv accelerometers placed on ankle wrist (*a* and *w* in Fig. 1) were used in order to recognise and measure activity levels of the participants. The GENEActiv accelerometers provided triaxial (*x,y,z*) acceleration measurements (± 8 g) with a sampling rate of 83 Hz. In order to create a recognisable pattern in data for synchronisation, the participants started the sequence of activities with a light jumping for 20 s while waving arms. The recorded signal of ankle and wrist per axis is presented in Fig. 2.

Indirect calorimetry The volume of oxygen (VO_2) and carbon dioxide (VCO_2) was measured per breath continuously during the activities, with a short break between the indoor and outdoor part of the protocol. The calorimetry measurements were obtained through the COSMED K4b^2 (McLaughlin et al. 2001) device, with a portable unit on the torso and a flexible mask covering the participant's nose and mouth (*K*4 in Fig. 1). The mask is connected to the portable unit that contains O_2 and CO_2 analysers, a sampling pump, a barometric sensors and electronics. The gas analyser measures the exchange of oxygen and carbon oxygen (in $\mathrm{ml\,kg^{-1}}$) and outputs PAEE metrics such as energy expended per minute, *EEm* in Kcal per minute, or per hour, *EEh* in Kcal per hour or *MET*, where 1 MET at rest equals 1 Kcal/kg/h. Measurements in these three units can be straightforwardly translated between one another. The COSMED

Table 3 Description of the final study population and their average COSMED measurements

	Indoors (12 Activities)	Outdoors* (4 Activities)	Total (16 Activities)
N (female %)	13 (27%)	18 (46%)	31 (35%)
Age in years (SD)	66.8 (4.5)	64.9 (4.4)	65.7 (5.0)
Height in cm (SD)	172.1 (8.3)	176.2 (7.4)	174.5 (7.9)
Weight in kg (SD)	82.2 (13.5)	83.7 (10.2)	83.1 (11.5)
BMI in kg/m^2 (SD)	27.7 (3.5)	26.9 (2.0)	27.2 (2.7)
EEm in Kcal (SD)	2.9 (0.4)	6.1 (0.9)	3.8 (1.1)
BR in breaths per sec (SD)	0.30 (0.05)	0.39 (0.04)	0.31 (0.04)

*4 out of 18 participants with outdoor activities did not perform cycling (1 female)

metrics are calculated per breath based on formula that combines VO_2 and VCO_2 measurements and is similar[5] to the Weir formula (Weir 1949):

$$\text{Metabolic rate (calories per minute) or EEm } = 3.94\,VO_2 + 1.11\,VCO_2$$

The output from this sensor in *EEm*, see Fig. 2, was used as our target for training and evaluating our PAEE estimation models. The sampling rate (SR) of the target is equal to the breathing rate of the participant and depends also from the activity at a specific moment. This results in an SR that is not stable, with a mean SR among all existing data being equal to 0.3 Hz.

Before every individual started the sequence of activities, the system was manually calibrated according to the manufacturer instructions. If a device was severely limiting a participant's movement (COSMED unit and battery weighs 1.5 kg), it was removed and the participant was excluded from our current analysis.

3.4 Resulting dataset

There were 35 participants recruited in the GOTOV dataset, from whom 31 participants had both COSMED (indirect calorimetry) and GENEActiv (ankle, wrist accelerometer) data. Of those, there were 13 participants with only indoor activity data, so 12 out 16 activities. Finally, for all the other participants with both indoor and outdoor activities, there were 4 participants that did not perform the outdoor cycling activity.

Table 3 presents the participant-level data of this study and the average measurements of COSMED. In detail, in the first block it displays the number of female participants out of the total 31 participants, and the average (mean and SD) age, height, weight and BMI. Furthermore, we can see the average EEm measurements by COSMED and breathing rate (sampling rate) for indoors, outdoors and total. From that, it is observed that there is a clear difference between the indoors and outdoors measurement in terms of EEm, where the mean outdoor EEm measurement is a bit more than double that of the indoor. This is something expected since the outdoor mea-

[5] COSMED uses a slightly different estimation (unknown formula) giving an average overestimation of approx. 0.0098976 cal compared to Weir.

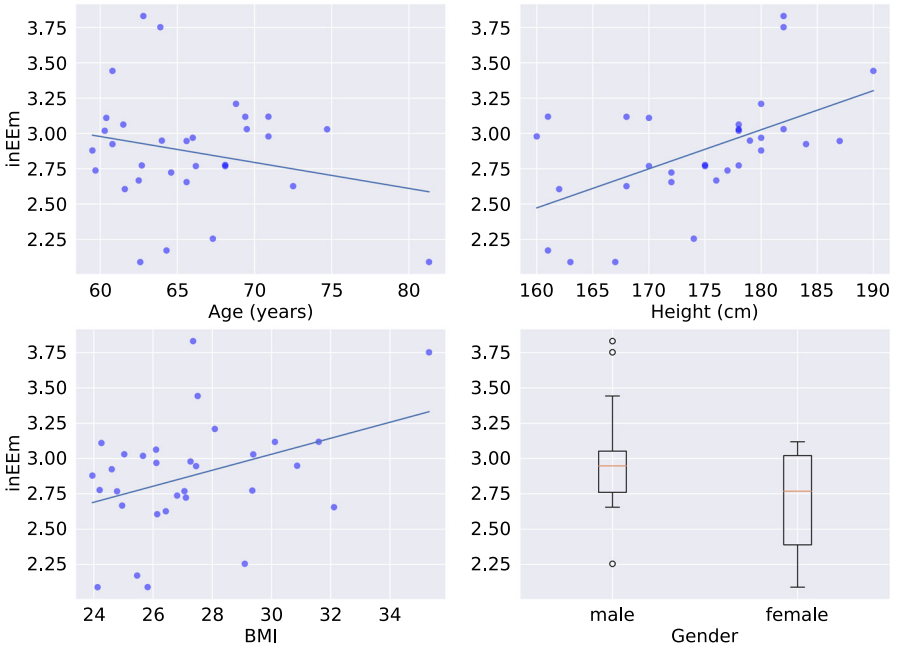

Fig. 3 Trend of indoor activities Energy Expenditure (y-axis) across Age, Height, BMI and Gender

surements include high intensity activities such as walking and cycling with a bigger range of EEm values compared to the indoors that have a smaller range. Similarly, the breathing rate is higher for the outdoor activities, which implies more data inputs for the same window of time when compared to the indoors (outdoors EEm SR higher than indoors), again as expected.

In total, the data set includes 2.8 hours of sedentary activity (MET $<$ 1.5), 5.4 hours of light activity (1.5 $\leq$ MET $<$ 4), 1.8 hours of moderate (4 $\leq$ MET $<$ 6) and 0.73 hours of vigorous activity (6 $\leq$ MET). An initial view of the dataset is presented in Fig. 3, where the indoors energy expenditure measurements is plotted against gender, age, height and body composition per participant. We plotted the indoors EEm since all 31 participants had indoors COSMED data. From the plots, we see that the trends from the GOTOV dataset confirm what is known from the literature. In detail:

- EE decreases with age (Frisard et al. 2007; Roberts and Dallal 2005).
- EE increases with height (Hills et al. 2014).
- EE increases with body composition (BMI) (Weinsier et al. 1992).
- EE in males is on average higher compared to the female participants (Keys et al. 1973).

4 Methodology

In this section, we explain the methodological contributions of the paper. In detail, we describe our model choice and its architecture. Following that, we analyse the steps

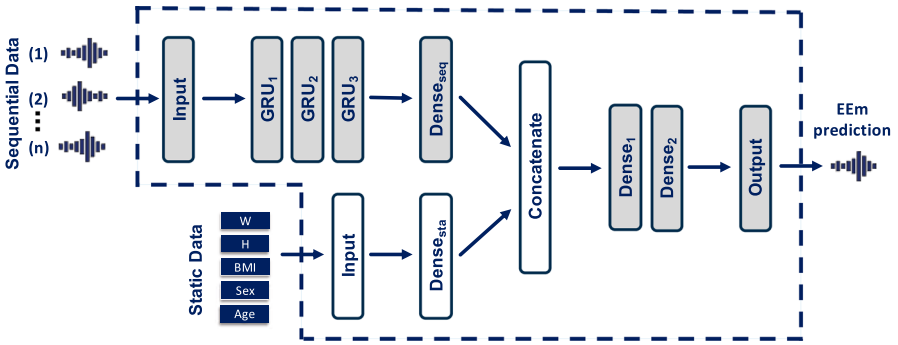

Fig. 4 Proposed model architecture combining both temporal and static data. The grey layers are sufficient when only temporal data are used (no static data)

of data preparation and their different combinations. Then, the training and evaluation process is explained. Finally, we summarise the experimental setup.

4.1 Modeling architecture

A Recurrent neural network (RNN) is a type of artificial neural network that has the ability to 'remember' older information from sequences. In more detail, an RNN contains feedback loops within its hidden layers whose activation at each time depends on that of the previous layer (Chung et al. 2015). Consequently, RNNs have a modeling advantage when used on sequential or temporal data over traditional ANNs. RNNs have been used for a variety of tasks, both regression and classification, such as natural language processing (Li and Xu 2018), speech recognition (Lee et al. 2018), in clinical application (Tomašev et al. 2019), and more recently, activity recognition from accelerometer data (Edel and Köppe 2016; Guan and Plötz 2017) and modeling of long-term human activity (Kim et al. 2017). Because PAEE is influenced by past activities (lag effect), RNNs could be a suitable modeling candidate for tackling the challenge of PAEE estimation.

Traditional RNN networks are known to struggle with information from long sequences due to the so-called *vanishing gradient problem* (Hochreiter 1998). The most popular solution to this problem is introducing *Long Short Term Memory* (LSTM) or Gated Recurrent Unit (GRU) layers. LSTM and GRU layers contain cells that act either as memory or gates controlling the information flow to the next layers. LSTMs contain 3 gates, namely, *forget, input* and *output*, while GRUs have only 2 gates, the *reset* and *update* gate. The reset gate controls which of the memory cell information needs to be forgotten. The update gate controls which information needs to be updated. This allows them to remember long sequences of information without losing relevant information. Adding to that, in a recent work of ours (Okai et al. 2019), we tested different RNN architectures with either LSTM or GRU layers for the task of activity recognition with predictors missing data. Based on these experiments we proved that RNNs with GRU layers and a proper architecture are robust for such a task.

Table 4 Proposed model architecture and training parameters

Input 1: Sequential data			
Layer	Nodes	Kernel regularizer (L2)	Dropout
GRU_1	32	0.0001	0.5
GRU_2	256	0.0001	0.5
GRU_3	32	0.0001	0.5
Layer	Nodes	Activation function	Dropout layer
$Dense_{seq}$	32	ReLU	–
Input 2: static data			
Layer	Nodes	Activation function	Dropout layer
$Dense_{sta}$	32	ReLU	–
Concatenate inputs			
Layer	Nodes	Activation function	Dropout layer
$Dense_1$	32	ReLU	0.2
$Dense_2$	16	ReLU	0.2
Model output: linear			
Training parameters			
Batch size	Epochs	Optimizer	Opimizing variable
512	50	Adam	Mean squared error (MSE)

As a result, our proposed RNN architecture is based on our recent work (Okai et al. 2019) and consists of an input layer followed by 3 GRU layers with 32, 256 and 32 nodes respectively, 2 dense layers with 32 and 16 nodes and an output layer (see Fig. 4 grey layers). Models are trained to minimize the mean squared error (MSE) using the optimization method Adam (Kingma and Ba 2015). To prevent over-fitting, a dropout ratio of 0.5 (50%) is applied to all three GRU layers. In Table 4, we describe in detail the different parameters chosen.

In order to test if participant-level data could improve PAEE estimation, we concatenated the aforementioned RNN setup with a single feedforward network, into the final architecture demonstrated in Fig. 4. The reason behind such an architecture is the need to model two types of data, temporal (sensor measurements, activities) and static data (participant-level). Therefore, we feed the accelerometer sequences to the GRU layers and at the same time, we feed the static data to a feedforward network. This feedforward network consists of an input layer and a hidden layer with 32 neurons. The output layers of both networks were concatenated and connected to 2 dense layers consisting of 32, and 16 neurons respectively, see Table 4. Finally, the output layer is made up of only a single neuron, which is used to predict the COSMED EEm values.

4.2 Data preparation and choices

In order to build PAEE estimation models using RNNs, there is a need for several transformations both in the predictors data (accelerometers, activities, participant-level data) and the target (COSMED EEm). As a first step, the target and numeric predictor data were z-normalized to have zero mean and a standard deviation of 1. Additionally, in order to model discrete predictors, like gender or activity class, label encoding was used (with values in $[0, n]$ with n being the number of values).

4.2.1 Indirect calorimetry as target data

COSMED produces energy expenditure measurements per breath, meaning that the target doesn't have a fixed sampling rate. On average, the COSMED sampling rate is 0.3 Hz, which means one input approximately every 3.3 s, see *EEm* signal in Fig. 2. In order to stabilize the sampling rate for the training, we down-sampled the COSMED signal to 0.1 Hz by taking the mean of every interval of 10 s. This way, we avoid the creation of more training data in periods with higher breathing rate, but we also smooth the outlier EEm values occasionally produced by COSMED. Finally, we assign a sequence of predictors per EEm value which captures the movements that preceded the EEm measurement (see box C of Fig. 5).

4.2.2 Predictors data to sequences

To train the RNN model, we need to build sequences, where each is associated with one EEm measurement. A sequence is defined as a finite list of inputs arranged in a definite order (Volchan 2002). For our problem, the order of inputs is based on time. The sequences represent the predictors data in the time immediately before every EEm measurement in windows of time, with specific number of inputs and resolution (sampling rate).

We have two types of inputs in our network. First, the activity data is of temporal nature, notably the accelerometer data include three numeric time series inputs for every device (see ankle, wrist in Fig. 2) and the activity labels (a discrete sequence). Second, we have the participant-level data that includes demographic information (age, gender) and body composition information (height, weight, BMI), as static attribute-value data. Figure 5 shows the different elements and steps taken to transform this data to training sequences. In detail, I, II, III, and IV in Fig. 5 display the different types of input, while for the accelerometer data (II), we display the extra steps needed in order to transform the signal to training sequences. In the following paragraphs, we explain the different sequence configurations developed and tested.

Accelerometers In order to transform the accelerometer signal into training sequences, see light shading in box A, Fig. 5, we need to decide on the number of inputs to be used (sequence size of the RNN), the length of the window that those will represent (time interval) and the resolution (sampling rate). These 3 variables are inter-connected, as presented in the following equation:

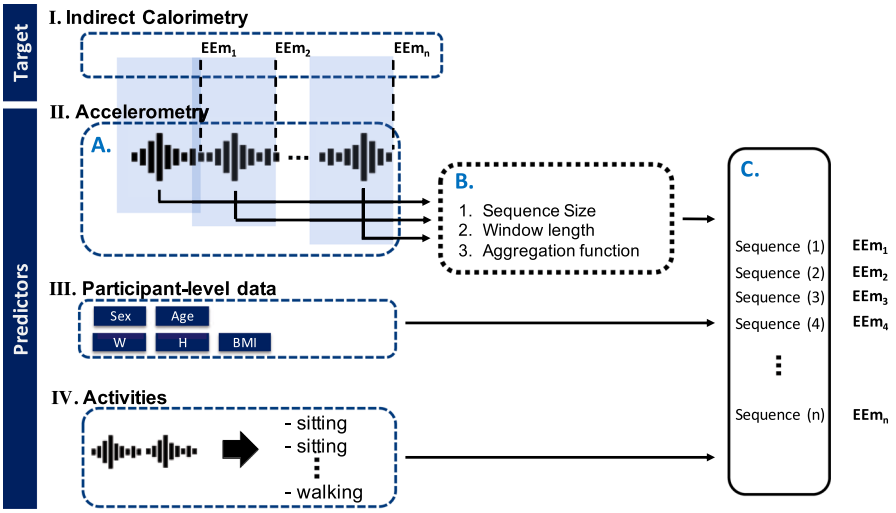

Fig. 5 Building sequences for temporal data

$$SR = \frac{Sequence\ Size}{Window\ Size},$$

where *Window Size* is calculated in seconds. For example, if we want a sequence with a size of 480 inputs to represent a time window of 240 s (4 min), we will need to down-sample the accelerometer data from 83 Hz (original SR) to 2 Hz, since the sampling rate (SR) depends on both sequence and window size.

On the one hand, longer sequences allow for higher sampling rates in the accelerometer data, but they will produce longer training times. On the other hand, for a fixed sequence size, a choice of longer time windows will result in lower sampling rate. Nevertheless, the choice of window length is crucial since long enough windows are needed in order to include any PAEE bias from activities performed further in the past. We experimented with different sequence sizes representing different intervals of time (time windows) and data resolutions. The down-sampling decisions are displayed in box *B* of Fig. 5.

In order to adjust the predictors to the given sequence sizes and window lengths, we need to down-sample the accelerometer data to the desired SR (*B*, Fig. 5). For this down-sampling, which aggregates several values into a single one, we compared two different aggregation approaches, one that uses the mean function, and one that makes use of statistical dispersion functions (standard deviation, interquartile range, percentiles difference).

Our motivation to use statistical dispersion measures comes from the fact that PAEE depends on the range of movement and therefore dispersion measures are more suitable for this task than the mean. As an example, see Fig. 6 where aggregated accelerometer values with mean and standard deviation (SD) are compared. Here, we can observe that during walking, the 3 axes of the ankle signal aggregated by SD (*Ankle SD* in figure) correlate nicely with the values of EEm compared to the *Ankle Mean* signal

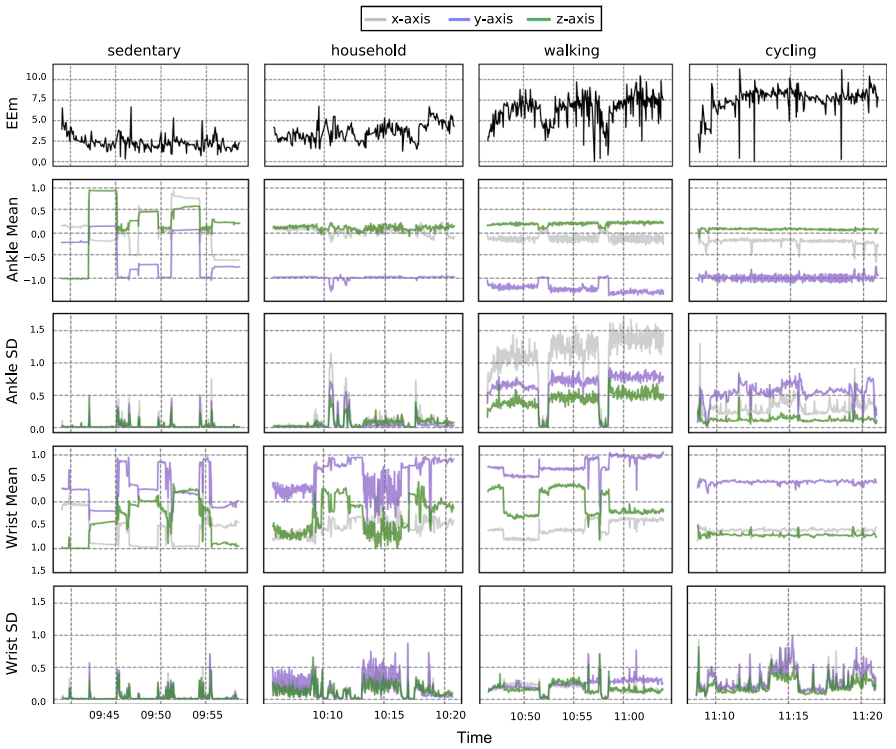

Fig. 6 The ankle and wrist accelerometer signal down-sampled to 1 input every 2.4 s (our optimal down-sampling) with *mean* and *SD*, compared to the recording of COSMED

which is represented in a more linear way. Similarly, during household activities, the *Wrist SD* correlates with EEm much more than *Wrist Mean* does.

As a simple example, let's consider 2 different movements represented in 2 windows of accelerometer data, W_1 and W_2. Let the windows also have different ranges of movement represented by only 2 values, $x_1 \in \{-2, 2\}$ and $x_2 \in \{-4, 4\}$. As a result, the energy spent for the movement in W_1 would be lower than in W_2 since the effort needed to go from -2 g to 2 g is lower than the effort needed to go from -4 g to 4 g. However, the mean magnitude in both windows is equal ($\text{mean}_{w_1} = \text{mean}_{w_2} = 0$). On the other hand, the standard deviation of these ranges are different, $\text{SD}_{w_1} = 2$ and $\text{SD}_{w_2} = 4$, which correctly captures of the relative expected energy spent per window. Concluding, the signal aggregated with statistical dispersion functions is more likely to capture the variation of PAEE compared to the ones of mean. For this reason, in this work we want to test this hypothesis (see Sect. 5).

Participant-level data Combined with the accelerometer data, we test whether participant-level data like demographics (age, gender) and anthropometric features (height, weight and BMI), see Table 3, could contribute to PAEE estimation (IV, Fig. 5). For this reason, we had to prepare such data input and combine it into the data sequences. The anthropometric data were z-normalized to have zero mean and

Table 5 Table of time spent and EEm (Kcal/min) per activity

Activity	Time (h)	EEm (SD)
Lying down	0.8 (7.4%)	2.4 (1.1)
Sitting	1.6 (14.8%)	2.0 (1.0)
Standing	2.5 (23.2%)	3.2 (1.7)
Household	2.2 (20.4%)	3.5 (1.6)
Walking	2.3 (21.4%)	5.2 (2.1)
Cycling	1.3 (12.1%)	8.2 (3.0)
Jumping	0.1 (0.6%)	3.1 (1.7)

a standard deviation of 1 and the gender was hard encoded. This way the model will take as an input a sequence of accelerometer data and the details of the corresponding participant.

Activity classes data Finally, we would like to test whether adding symbolic data in the form of a label describing current or past activities can be beneficial to estimate PAEE (*IV*, Fig. 5), when combined with either accelerometer data only, or with both accelerometer and participant-level data, as seen before in Bonomi et al. (2009) and Altini et al. (2015). In order to obtain such activity labels, we had to predict the activity types using learned activity recognition methods. We need to derive the labels from the acceleration data, since they will not be available in a free-living scenario either.

For this goal, we used a previously developed and published method that was already tested with the GOTOV devices (Paraschiakos et al. 2020). This model can produce activity predictions per second with an accuracy of more than 90% based solely on ankle and wrist accelerometers for 7-class activity classification. Through this model, we can predict the following 7 classes: *lying down, sitting, standing, household, walking, cycling* and *jumping*. Having the activity labels predicted per second, we encoded them and combined them with the accelerometer sequences as input to our model. Table 5 summarises the predicted classes and presents also some statistics about their EEm cost.

4.3 Training and evaluation

We trained and tested our models using *Leave One Subject Out* cross validation (LOSO-CV). This means that we train using all subjects (participants), leaving the data of one subject out as a test set. We then iterate the process in order to test all subjects separately. The aim of this type of cross-validation is that we emulate the future situation where we would like to process as yet unseen subjects. The LOSO-CV process prevents training set leakage within a subject, as normal cross-validation procedures might allow. Additionally, during training, 2 participants were selected as validation set, one with only indoor activities and one with all activities. These 2 sets were randomly chosen per subject and were the same across the different model settings tested in order to have fair comparisons. All models were trained for 50 epochs with a batch size of 512. After testing 4 different batch sizes (64, 128, 256, 512) we

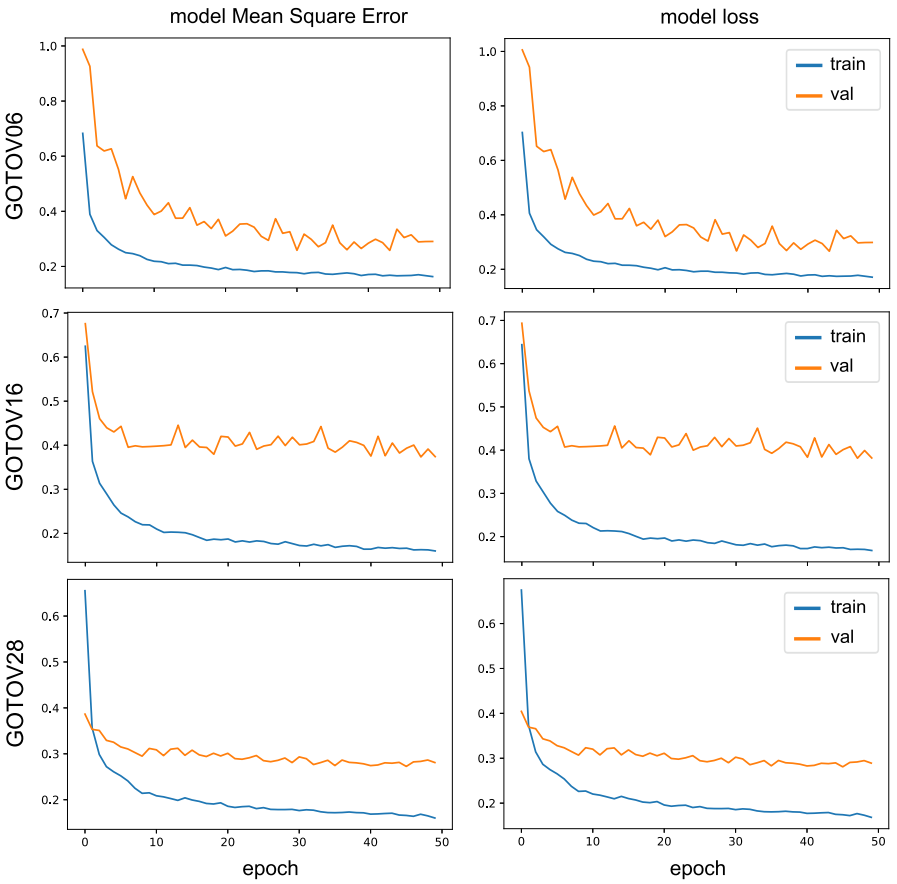

Fig. 7 Three examples of MSE error (left) and loss (right) during training for 50 epochs. The orange lines represent the evolution of MSE and loss for the training set and the blue for the validation set accordingly

selected the max (512) since the accuracy gain for the smaller ones was too little compared to the cost of training time. Similarly, we tested three sizes of training epochs (50, 100, 200) and it was proven that, with our optimal set-up of SD-aggregated signal, 50 epochs were enough for the model to converge while the cost of extra training epochs was not translated to significant performance gains. This can also be seen in Fig. 7 where the mean squared error (MSE) and loss evolution during training of 3 LOSO examples is presented.

We would also like to point out that the model's validation and test sets during LOSO-CV are used with their original sampling rate (once per breath). This means that we *trained* our models using the smoothed EEm values with a stable SR (0.1 Hz), as indicated in the previous section, but we *evaluate* them per breath (COSMED recordings). This way, we can see which model can fit better the input data since we evaluate our models by measuring their performance on the original EEm values, including the extreme COSMED measurements.

Hence, to get the overall performance of a model, we train 31 different models using LOSO-CV and we report their aggregated (median) result, as *Root Mean Squared Error* (RMSE) and *R-squared* (R^2). Additionally, we compute the RMSE and R^2 separately for indoor and outdoor data. There are multiple reasons behind this decision. First, since there are participants without outdoor data and there is a clear difference between EEm levels with the indoor ones (see Tables 3, 5), we can see how our models behave on low and high-intensity activities separately. Additionally, it is suggested in the literature (van Hees et al. 2009) that PAEE estimation of sedentary or low-intensity physical activities (typically performed indoors) is still a challenging task since their differences in acceleration magnitude are minor. Finally, our main focus is to estimate PAEE of older individuals and it is observed that this group of people spends significantly more time in sedentary or low-intensity activities (van Ballegooijen et al. 2019).

4.4 Experimental pipeline

In this section, we explain the experiments performed, the motivation behind their set-up and their order.[6]

Optimise data input First, we tested the architecture with only accelerometer data (grey in Fig. 4) comparing the different accelerometer aggregation functions into time windows of different sequence size and resolution (SR). The aggregation functions tested were *mean, standard deviation (SD), interquartile range (IQR)*, and *difference between 5th and 95th percentile (PD)*, with:

- sequence sizes of 10, 50, 160 and 480 inputs per sequence, and
- window lengths, for each sequence size, of 1, 2, 4, and 8 min.

Here, in order to avoid training and testing $4 \times 4 \times 4 = 64$ different combinations of sequence size, window length, and aggregation function, we optimize the search process by first comparing all aggregation functions with the longest sequence size (480) representing the 4-min windows. Since we observed that the sequences built with statistical dispersion functions had very similar performance, we subsequently fix our dispersion measure to *SD*, and compared this with *Mean* for the remaining combinations (2 functions $\times 9$ combinations $=$ 18 in total). In the end, when the optimal setting of window and sequence size is found for the *SD* and *Mean*, we tested them also for the *IQR* and *PD* aggregations. The training parameters used were a *batch size* of 512, for 50 *epochs* for all experiments since we observed that with this combination the model converged faster.

Anthropometrics and activity classes data As a second analysis step, we tested whether the addition of participants-level data or the predicted activity classes improves the performance. In order to do that, we make use of the complete architecture and parameters presented in Fig. 4 and Table 4 and the best combination of aggregation function (SD), window size ($w = 2$ min) and sequence size (N = 50), as concluded from the step above. In the end, we compare the performance of this model set-up using: (1)

[6] To make our experiments, scripts and results accessible to other researchers (open-source) we share them on the following git repository, https://github.com/parastelios/GOTOV_PAEE_publication.

Table 6 Table of architectures tested at ablation study

Architecture	GRU layers	Dense layers	Data
5xGRU_3xDense	1, 1, 2, 3, 3	Seq, sta, 1, 2	a, w
4xGRU_3xDense	1, 1, 2, 3	Seq, sta, 1, 2	a, w
3xGRU_3xDense	1, 2, 3	Seq, sta, 1, 2	a, w
3xGRU_2xDense	1, 2, 3	Seq, sta, 1	a, w
2xGRU_3xDense	1, 2	Seq, sta, 1, 2	a, w
2xGRU_2xDense	1, 2	Seq, sta, 1	a, w
1xGRU_2xDense	1	Seq, sta, 1	a, w
3xGRU_3xDense_a	1, 2, 3	Seq, sta, 1, 2	a
3xGRU_3xDense_w	1, 2, 3	Seq, sta, 1, 2	w
1xGRU_2xDense_a	1	Seq, sta, 1	a
1xGRU_2xDense_w	1	Seq, sta, 1	w

The number and descriptions of the layers refer to the ones in Fig. 4 and Table 4. Data refers to ankle (a) and/or wrist (w) inputs used for training

accelerometer and participants level data (GRU_{ID}) and (2) accelerometer, participant-level and activity classes data (GRU_{ID_AC}).

Ablation study Subsequently, we performed an ablation study for our proposed architecture, and we tested our model against different RNN layers. For these experiments we used the best combination of data resolution and data combination as found in the previous steps. For the ablation study, we tested 2 more coomplicated and 3 simpler architectures of our model by excluding different layers of our model. The different architectures are presented in Table 6. Additionally, we compared the performances of our proposed model (3xGRU_3xDense) to the minimal one (1xGRU_2xDense) using data (train and test) from ankle or wrist alone. The motivation behind this is to test how robust each architecture is to training data from only ankle or only wrist since in our previous work (Okai et al. 2019), the maximum (3xGRU_3xDense), was proven to be robust to missing data for a classification task (activity recognition) and that is why it was selected. Furthermore, most of PAEE related work tests models that exploit data from single devices and we would like to compare our model to theirs (see Sect. 2). Finally, we tested our maximal architecture with two different types of RNN layers, the long short term memory (LSTM) and simple RNN layers. For all experiments, we compared the performance using both *R-squared* and *RMSE* for all activities, and for indoor and outdoor ones separately.

Compare to related work Following to ablation study, we compared our proposed model against state-of-the-art methods as presented in Sect. 2. We trained a linear model (LM), a linear mixed model (LMM), a random forest regressor (RFr), an artificial neural network (ANN) and a convolutional neural network (CNN) using our dataset. In detail, we tested the LM, LMM and ANN as presented in Montoye et al. (2017a) with the proposed set of 30 time-domain features for non-overlapping windows of 30 s per device. In order to have comparable results, we used the same set of features to train an RFr similar to the one that Ellis et al. (2014) introduced but

using 1000 trees (compared to 500) as O'Driscoll et al. (2020) recently presented to be more robust. Finally, we trained a CNN as presented in Zhu et al. (2015) using both the authors' data resolution with a window of 5.12 s for sequences of size 256 inputs, and our approach of down-sampled predictors with standard deviation for a window of 2 min and a sequence size of 50 inputs. About that, since the authors did not include the number of epochs and batch size, we decided to train the CNN for 50 epochs and a batch size of 512, similar to our GRU approach.

Here, we need to clarify that all of the above publications (Montoye et al. 2017a; Ellis et al. 2014; Zhu et al. 2015) recommend models with data from a single accelerometer placed at different body locations per study. Even when multiple devices exist, only (Ellis et al. 2014) combined them and tested their performance against the single ones. In contrast, our purpose in this work is to predict PAEE by combining the data of ankle and wrist and not compare them separately. However, in order to compare our work to the above methods, we also trained our optimal architecture with only ankle or wrist data.

Demonstrate the model's use in the future Finally, we test our selected model's performance over different EEm aggregation windows and activity types. The motivation behind this is twofold. First, we wanted to test how our model performs in a free-living setting where breathing rate is not included, and second, in order to have comparable results to the literature where models are evaluated in aggregated windows of 30 s or 1 min. In particular, we report the performance across different EEm windows from the original COSMED sampling rate (breath by breath), to 10, 30 s and 1, 5, 60 min aggregations and per activity type. This way, we can have an idea of how our approach can be used to estimate PAEE of longer windows.

5 Results and discussion

In this section, we present and discuss the results of the experiments performed. First, we discuss the training input optimisation (Sects. 5.1, 5.2), followed by the results of the ablation study (Sects. 5.3, 5.4) and the comparison of our results to related work (Sect. 5.5). Finally, we demonstrate the results of our suggested model for different windows of PAEE aggregations and activities (Sect. 5.6).

5.1 Standard deviation as optimal aggregation function

Here, we present the performance of the different input data setups. Since it is not feasible to display all 64 combinations tested, Table 7 displays the best setup per aggregation function. In the first column, the aggregation function is displayed, followed by its resulting best data setup (sequence size, window length, sampling rate distribution). Then we compare their R^2 and RMSE in total, indoor, and outdoor activities. Additionally, the last column indicates the significance of the difference between mean and the rest functions in terms of R^2, using a paired t-test.

Examining Table 7, we observe that models built with statistical dispersion functions outperform significantly the one using the mean, for $\alpha = 0.05$, all p-values are

Table 7 Comparing the performance of different data setups

Model		R^2	inR^2	outR^2	RMSE	inRMSE	outRMSE	p-Value
Mean	SeqSize = 480 WinSize = 4 min SR = 2 Hz	0.38	0.23	0.38	1.46	1.24	2.32	–
SD	SeqSize = 50 WinSize = 2 min SR = 0.42 Hz	0.45	0.31	0.33	1.35	1.16	2.03	0.01*
IQR	SeqSize = 50 WinSize = 2 min SR = 0.42 Hz	0.41	0.29	0.33	1.39	1.18	2.15	0.02*
PD	SeqSize = 50 WinSize = 2 min SR = 0.42 Hz	0.43	0.30	0.28	1.36	1.17	2.03	0.01*

The p-value corresponds to the paired t-test of mean with SD, IQR, in terms of R^2, with * pointing out when $p < 0.05$

significant. Additionally, when we feed our model with SD-aggregated accelerometer data instead of averaging, not only is the model's performance significantly improved, but this improvement is achieved by using approximately 10 times less input data (sequences of 50 versus 480 inputs) and double the window size (4 versus 2 min), leading to approximately 10 times lower training time. This is because statistical dispersion metrics can represent the original signal in a more characteristic way compared to averaging with mean (Lyden et al. 2011).

In detail, for the estimation of PAEE using the developed RNN, a two-minute window can be represented by statistical dispersion functions using only 50 inputs per sequence without significant loss of accuracy compared to windows down-sampled by averaging. This is of importance when applying the RNN model to data sets with large sample size, e.g. intervention studies (hundreds of participants or more) or epidemiological studies like UK Biobank (thousands) (Sudlow et al. 2015). Accurate estimation of PAEE in such studies will be relevant so that personal advise can be given to older persons with respect to the most effective and still achievable beneficial changes in lifestyle. This sums up to a computationally efficient and accurate method to estimate PAEE in older adults.

Furthermore, comparing the models built with SD, IQR and PD, there is no clear performance difference. That is because all three measures are similar in behaviour, and their main differences are in the magnitude of training values. Based on that, if we have to choose one of them, we believe that the SD model seems to be slightly better than the others, both in terms of R^2 and RMSE. Adding to that, SD is more intuitive as a metric compared to IQR and PD. Therefore, for the rest of our analysis, we will focus on the model built with the following settings for accelerometer data: (1) a sequence

Table 8 Comparing models with participant-level data and/or activity classes

	R^2	inR^2	outR^2	RMSE	inRMSE	outRMSE	p-Value
GRU	0.45	0.31	0.33	1.35	1.16	2.03	–
GRU_{ID}	0.55	0.41	0.36	1.25	1.09	2.05	0.02*
GRU_{AC}	0.42	0.32	0.35	1.33	1.14	1.96	0.38
GRU_{ID_AC}	0.50	0.33	0.40	1.29	1.13	2.00	0.30

The p-value corresponds to the paired t-test of GRU (only accelerometry data) with any data addition, GRU_{ID} (accelerometer and participant-level data), GRU_{AC} (accelerometer and activity class data), and GRU_{ID_AC} (accelerometer, participant-level and activity classes), in terms of R^2, with * indicating when $p < 0.05$

size of 50 inputs, (2) representing a time window of 2 min, (3) down-sampled to a resolution of $SR = 0.42$ Hz with SD.

5.2 Adding participant-level data results in better PAEE estimation

Table 8 demonstrates the effect of participant-level data and activity classes in our RNN model to estimate PAEE (first row marked GRU). The addition of participant level data (such us age, sex, height, weight, and BMI) improves the results, both in terms of R^2 and RMSE error (GRU_{ID} model). In more detail, the model's performance improves significantly (p = 0.02) from 0.45 (GRU) to 0.55 (GRU_{ID}) for R^2, while for RMSE the error decreased from 1.35 Kcal/min to 1.25. This development is mainly a result of the improved performance in the lower intensity activities (indoor activities), where RMSE decreased from 1.16 to 1.09 Kcal/min and R^2 from 0.31 to 0.41. This is quite an important observation since it is mentioned in the literature that estimating PAEE of lower intensity activities is challenging (van Hees et al. 2009) and it seems that anthropometric data can help in this respect.

On the other hand, adding activity labels doesn't seem to improve the results (see GRU_{AC} in Table 8), where the R^2 drops (not significantly though), and RMSE increases. Interestingly, the addition of (predicted) activity classes, even when combined with participant-level and accelerometer data (model GRU_{ID_AC}), did not produce any significant improvement to our model ($R^2 = 0.50$, ($p = 0.3$)). This is a notable observation for our architecture since it contradicts with what is shown in previous work (Bonomi et al. 2009; Altini et al. 2015). It seems that the way RNNs model the input sequences and its ability to 'remember' past information, the exact activity labels are not needed for efficient PAEE estimations. Therefore, when the objective is only PAEE estimation and not its association with specific activities, there is no need for applying activity recognition algorithms beforehand. Still, we need to mention that our dataset might not be ideal in order to prove this point, since activity windows are not equal and the breaks between each activity are not long enough to avoid the PAEE lag effect.

Table 9 Comparing the different architectures of the ablation study, where t is the average training time per run in seconds

Architecture	R^2	inR^2	outR^2	RMSE	inRMSE	outRMSE	t (s)
5xGRU_3xDense	0.51	0.38	0.41	1.38	1.12	1.87	732.8
4xGRU_3xDense	0.49	0.40	0.45	1.41	1.13	1.92	656.5
3xGRU_3xDense	0.55	0.41	0.36	1.25	1.06	2.05	607.4
3xGRU_2xDense	0.5	0.39	0.42	1.28	1.09	1.97	589.8
2xGRU_3xDense	0.47	0.35	0.39	1.41	1.15	1.92	237.2
2xGRU_2xDense	0.52	0.37	0.44	1.30	1.08	2.00	191.6
1xGRU_2xDense	0.52	0.39	0.46	1.32	1.09	1.99	98.2
3xGRU_3xDense_a	0.50	0.33	0.35	1.32	1.09	2.00	377.1
3xGRU_3xDense_w	0.41	0.23	0.29	1.37	1.25	2.17	352.1
1xGRU_2xDense_a	0.45	0.27	0.42	1.38	1.16	2.03	60.8
1xGRU_2xDense_w	0.25	-0.01	0.02	1.57	1.35	2.60	56.6

5.3 Ablation study

For the ablation study, we are comparing our proposed architecture against simplified architectures to determine whether the increased complexity is warranted. As Table 9 demonstrates, in terms of R^2 and RMSE, the proposed architecture is optimal ($R^2 = 0.55$, RMSE $= 1.25$), although reasonably similar results could be obtained with architectures with fewer layers, e.g. 2 GRU layers and 2 dense layers (*2xGRU_2xDense*). In more detail, separating indoor and outdoor activities, we note that simpler architectures with fewer GRU layers might perform better on *outdoor* activities, with the best results obtained by the simplest architecture tested *1xGRU_2xDense* with an $R^2 = 0.46$. This probably has to do with the fact that smaller networks are more robust against overfitting. On the other hand, for the challenging task of estimating PAEE of lower intensity activities (*indoor* activities), an extra GRU layer, as in the proposed model, is still the best option with an $R^2 = 0.41$. Still, adding 2 further GRU layers has no beneficial effect estimating indoor activities PAEE, see *5xGRU_3xDense*.

Subsequently, we study the benefits of two devices versus only a single device on the ankle or wrist. The results demonstrate that a moderate price is paid for removing one device from the proposed architecture [from $R^2 = 0.55$ to $R^2 = 0.50$ for *3xGRU_3xDense_a* (-9%), and $R^2 = 0.41$ for *3xGRU_3xDense_w* (-25%)], whereas for *1xGRU_2xDense*, removing a device incurs a large penalty: from $R^2 = 0.52$ to $R^2 = 0.45$ for *1xGRU_2xDense_a* (-13.5%), and $R^2 = 0.25$ for *1xGRU_2xDense_w* (-52%). We see the same phenomenon in RMSE and within the indoor and outdoor activities, showing that the proposed architecture is more robust against removal of a device.

All in all, we observe that our proposed architecture (bold in Table 9) has a small advantage over the others, so if optimal accuracy is a priority, this architecture is the model of choice. However, if efficiency is important, quite reasonable results can still be obtained with smaller architectures. We feel the decent accuracy across the board

Table 10 Comparing the different RNN layers

RNN layer	R^2	inR^2	outR^2	RMSE	inRMSE	outRMSE
GRU	0.55	0.41	0.36	1.25	1.06	2.05
LSTM	0.48	0.40	0.41	1.29	1.07	1.98
simpleRNN	0.13	− 0.35	− 0.05	2.02	1.66	2.62

is due to the rich representation of data (SD, 2-minute windows and anthropometric data). There appears to be no added benefit for larger networks than the proposed one, although larger models still work reasonably well. Moving on to the bottom half of Table 9, where data from a single device is used, the proposed architecture (top two rows) handles the loss of data better than a smaller architecture (bottom two rows). This advantage was expected since we selected this architecture (*3xGRU_3xDense*) based on previous work of ours (Okai et al. 2019), where it was proven to be robust for the task of activity recognition under missing data. Concluding, if training speed is the priority, the simpler architectures are still reasonably accurate, but if there is a need for a model to be robust to data loss and lower intensity activities (indoor activities), the proposed one is preferred.

5.4 Comparing different RNN layers

Adding to the ablation study, we replaced the GRU layers on our proposed architecture with 2 other RNN layers, the LSTM and simple RNN. The results in Table 10 demonstrate that interchanging the GRU layers with LSTM ones will not cost much of performance ($R^2 = 0.49$) compared to the ones of simple RNN that did not manage to fit at all ($R^2 = 0.12$). However, training the same model with LSTM layers compared to GRU is more costly since every epoch takes almost 3 times longer than with the GRUs. Based on that, using GRU layers, we manage to have a better performance with a lower computational cost.

5.5 Our proposed model versus the state-of-the-art

In order to fairly compare our best model (GRU_{ID}) to the ones presented in related work (see Sect. 2), we performed our analysis in two steps. First, we compared our architecture to the convolution architecture proposed by Zhu et al. (2015) and then to the models proposed by Montoye et al. (2017a) and Ellis et al. (2014) (LM, LMM, RFr, ANN). We divided our analysis this way since the CNNs are tested with the original EEm (breath rate) while the rest of the methods are tested with 30-s aggregations of the target.

In detail, we tested the CNN using two different data settings, one with windows of 5.12 s for a sequences of 256 inputs ($CNN_{5\,sec}$), as suggested by the authors, and one with our best data set-up of 2-min windows aggregated with SD and sequence size = 50 ($CNN_{2\,min}$). Table 11 demonstrates in terms of R^2 the performance per model and per devices. It is clear that the CNN developed with the short window of 5 s

Table 11 Comparing R^2 score of our proposed model (GRU_{ID}) to CNN_{5sec} for windows of 5 s (Zhu et al. 2015 set-up) and CNN_{2min} with 2-min window

Method	Ankle & wrist			Ankle			Wrist		
	R^2	inR^2	outR^2	R^2	inR^2	outR^2	R^2	inR^2	outR^2
GRU_{ID}	0.55	0.41	0.36	0.50	0.33	0.35	0.41	0.23	0.30
$CNN_{5\,sec}$	0.29	−0.12	0.21	0.20	−0.13	0.19	−0.05	−0.45	−0.08
$CNN_{2\,min}$	0.49	0.28	0.36	0.37	0.23	0.24	0.29	−0.19	−0.05

Table 12 Comparing R^2 score of our proposed model (GRU_{ID}) with RFr-random forest regressor (Ellis et al. 2014), LMM-linear mixed model, ANN-artificial neural network and LM-linear regression (Montoye et al. 2017a), using 30-s features

Method	Ankle & wrist			Ankle			Wrist		
	R^2	inR^2	outR^2	R^2	inR^2	outR^2	R^2	inR^2	outR^2
GRU_{ID}	0.72	0.60	0.56	0.65	0.53	0.62	0.55	0.40	0.49
RFr	0.55	0.30	0.43	0.57	0.23	0.43	0.33	0.03	0.17
LMM	0.43	−0.07	0.47	0.29	−0.26	0.41	0.24	−0.19	−0.22
ANN	0.37	0.06	0.27	0.50	0.12	0.45	0.20	−0.26	0.24
LM	0.35	−0.12	0.45	0.24	−0.37	0.41	0.22	−0.37	−0.12

($CNN_{5\,sec}$) fails to explain enough of the EEm variation for all set-ups. On the other hand, the $CNN_{2\,min}$ has a somewhat comparable performance to our set-up (GRU_{ID}) for ankle and wrist data combined, with an $R^2 = 0.49$ versus $R^2 = 0.55$. However, when we compare per single device, the GRU_{ID} set-up clearly outperforms the CNN with $R^2 = 0.50$ for ankle and $R^2 = 0.41$ for wrist, compared to the CNN with $R^2 = 0.37$ and $R^2 = 0.29$, for ankle and wrist respectively.

To conclude, when our approach is compared to another deep learning approach, we see that the performance of our model outperforms the one of the CNNs both in combined ankle and wrist, as well as when compared to single-device settings. Remarkably, we can observe here too the effect that our robust representation of data (similar to the ablation study) has on the performance of the tested CNN model. Furthermore, it is clear that using longer windows of training data, down-sampled with SD, gives a big advantage in explaining PAEE variance.

Next, Table 12 demonstrates the performance of the proposed model and the methods trained with handcrafted features and both tested with 30-s EEm aggregations. Here, we observe that our method clearly outperforms all others discussed above. In detail, the GRU_{ID} model manages to explain almost 72% of the total EEm variation (aggregated EEm) when both ankle and wrist predictors are used, while the second best (RFr) explains only 55%. When data from indoor activities only is used for testing, our model captures 60% of EEm variation which is double the second method (RFr) in performance. Similar to that, when data from ankle or wrist alone are used, our model explains more than 50% of aggregated EEm variation, with $R^2 = 0.65$ for ankle and $R^2 = 0.55$ for wrist.

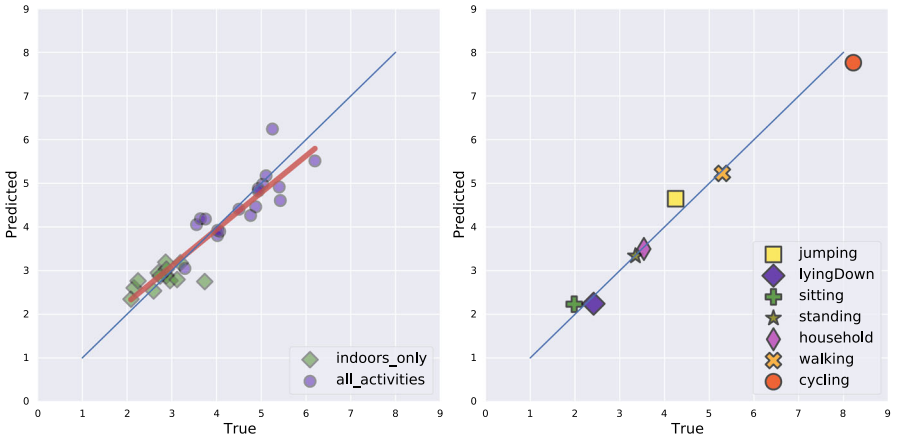

Fig. 8 Scatter plot of mean true over mean predicted EEm, per participant (left) and activity (right)

When comparing the rest of the methods, the random forest regressor (RFr) clearly outperforms LMM, ANN and LM with an $R^2 = 0.55$ for the ankle and wrist combined predictors and $R^2 = 0.57$ for the ankle only. Interestingly, ANN performs pretty well when only ankle predictors are used for training with an $R^2 = 0.50$, however this performance is mainly based to the higher intensity activities (outdoor walking, cycling) since in$R^2 = 0.12$ while out$R^2 = 0.45$. Still, when wrist and ankle are combined, ANN's performance drops to $R^2 = 0.27$. On the contrary, the LMM manages to capture quite well the variation of the outdoor EEm for both ankle and wrist predictors combined ($R^2 = 0.47$) and for ankle alone ($R^2 = 0.41$). Finally, we observe for all models that using wrist features alone is not enough to explain enough of the EEm variation.

Summarizing, comparing our model to methods that use a set of handcrafted features as predictors, it is clear that our approach outperforms the rest of the modeling methods. This is probably due to our approach being able to incorporate longer windows of predictors data. This way, EEm bias from a longer activity time is taken into account. In order to fairly compare our model with the other modeling choices, we had to test our model in aggregated windows of the target (30-s aggregations). Such aggregation has the effect of smoothing the per-breath PAEE measurements, creating a less noisy target. As a result, the over- or under-estimations are also smoothed out and the model performance seem improved. This model application is close to how we intend to use our model with free-living accelerometer data in the future. Therefore, in next section, Table 12 we also present the performance of our approach with similar scenarios of target aggregation.

5.6 Demonstration of GRU_{ID} model estimating PAEE

To appreciate the general performance of the GRU_{ID} model, consider Figs. 8 and 9. In Fig. 8, we see two scatter plots of the average true over average predicted EEm value per participant (left) and per activity class (right). In the left plot, the dots (in blue)

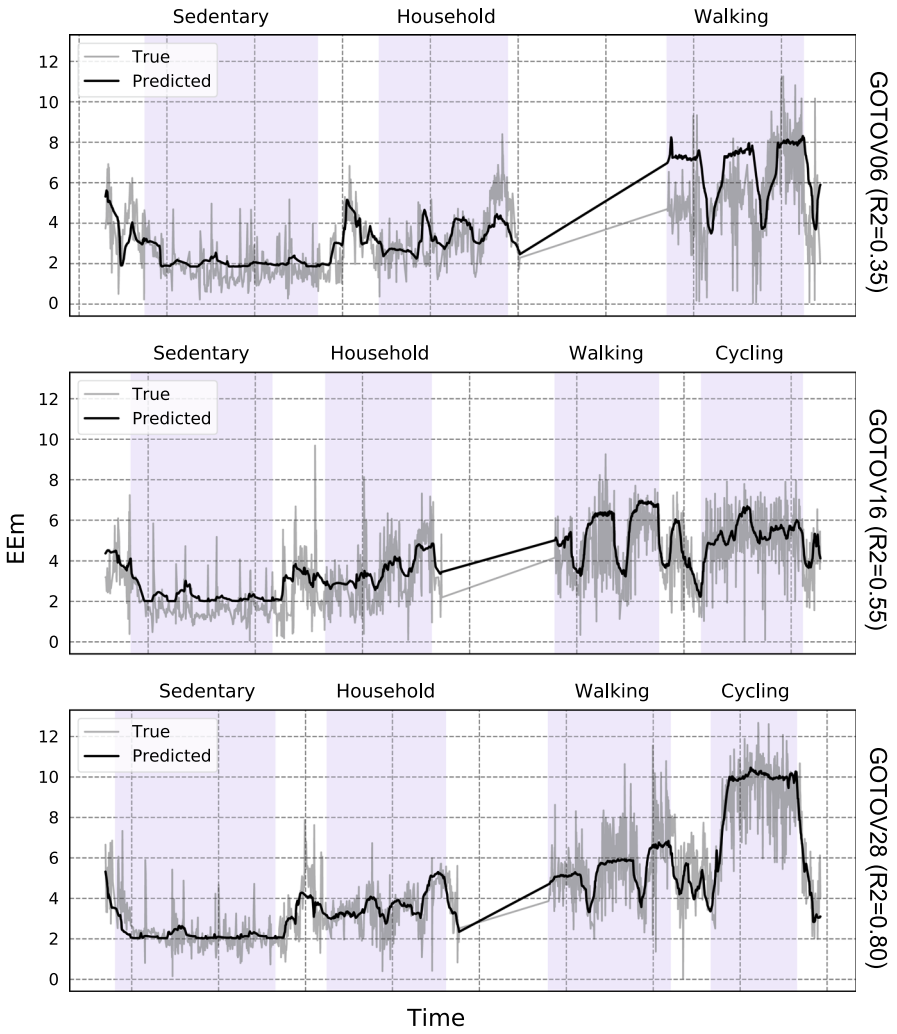

Fig. 9 True versus Predicted EEm per breath for participants with lowest (top), median (middle) and higher (bottom) R^2 examples, when indoor and outdoor activities are included

display the participants with both indoor and outdoor activities, while the diamonds (in green) represent participants with only indoors data. Adding to that, the trend line (in red) is used to compare the predictions to the main diagonal (blue, representing $x = y$) which is the ground truth. From that, we observe that our model has on average a good fit. However, it slightly overestimates the lowest EEm values (red line above main diagonal) and underestimates the highest ones (red line below main diagonal). In the right plot of Fig. 8, we can observe that our model captures really well the average EEm per activity since the average PAEE estimated for all the activities is either on or really close to the ground truth line. However, evaluating the performance over one activity class averages out a lot of the EEm signal.

Table 13 Comparing true and predicted EEm over different aggregations

Aggregation	R^2	inR^2	outR^2	RMSE	inRMSE	outRMSE
Per breath	0.55	0.41	0.36	1.25	1.09	2.04
10 s	0.65	0.51	0.50	0.95	0.89	1.58
30 s	0.72	0.60	0.56	0.82	0.74	1.40
1 min	0.78	0.67	0.62	0.76	0.64	1.34
5 min	0.82	0.78	0.63	0.62	0.48	0.97
60 min	0.86	–	–	0.40	–	–

Adding to that, in Fig. 9, we plotted the predicted over true PAEE (COSMED) values (recordings per-breath) for 3 participants that performed both indoor and outdoor activities and have the worst, median and best fit, $R^2 = 0.35$, $R^2 = 0.55$ and $R^2 = 0.80$ respectively. Here, we see that the model overall captures nicely the trend of the true EEm, as the black line (predicted EEm) follows the longer-term changes of the grey line (true EEm). However, the short-term behaviour of the target is not captured sufficiently, except for the sudden changes. We need to point out here that, while our models are tested on the real data (per breath), they were trained with averaged EEm values of 10 s both for smoothing out any non-generalizable noise (high peaks in grey in Fig. 9) of the data and for training with a stable sampling rate, since COSMED produces data per breath. As a result of this choice, we can see that our predicted EEm values do not capture the high-frequency fluctuations, but follow the average trend on a 10-second scale. Finally, the gap in the middle of the plots is the transition in the protocol from indoor to outdoor activities, where no COSMED measurements took place.

Subsequently, in Table 13, we present the performance of the GRU_{ID} model by different target aggregations. We aggregated the original and predicted EEm values at 10 and 30, and at 1, 5 and 60 min. This is really useful, since in a free-living setting, the model will be used to estimate PAEE for aggregated windows. From Table 13, it is observed that even with the shorter window of aggregation (10 s) the model's performance improves substantially both for R^2, from 0.55 to 0.65, and RMSE, from 1.25 to 0.95 Kcal/min. Additionally, for commonly used time frames of 30-second and 1-minute windows (Staudenmayer et al. 2009; Montoye et al. 2017a; Ellis et al. 2014; O'Driscoll et al. 2020), the model has an RMSE of only 0.86 and 0.76 Kcal/min respectively with the predictions explaining more than 70% of the original EEm signal variation for both aggregations, with $R^2 = 0.78$ for the 1 min.

Finally, we can observe that our model has different performance for indoor and outdoor activities. In detail, when comparing windows of 30 s, for indoor activities the RMSE is equal or less that 1.0 Kcal/min, while for outdoor activities, RMSE is much higher (walking RMSE = 1.20; cycling RMSE = 1.50), see Table 13. Characteristically, when we compare our model's performance per participant, we see a slight overestimation of EEm for those with only indoor activities (low-intensity activities) and a slight underestimation for those with high average EEm, see Fig. 8. Especially, when ordering by R^2 (see Fig. 10), we can clearly see that for participants with lower median EEm, the model explains less of its variance (lower R^2), while for participants

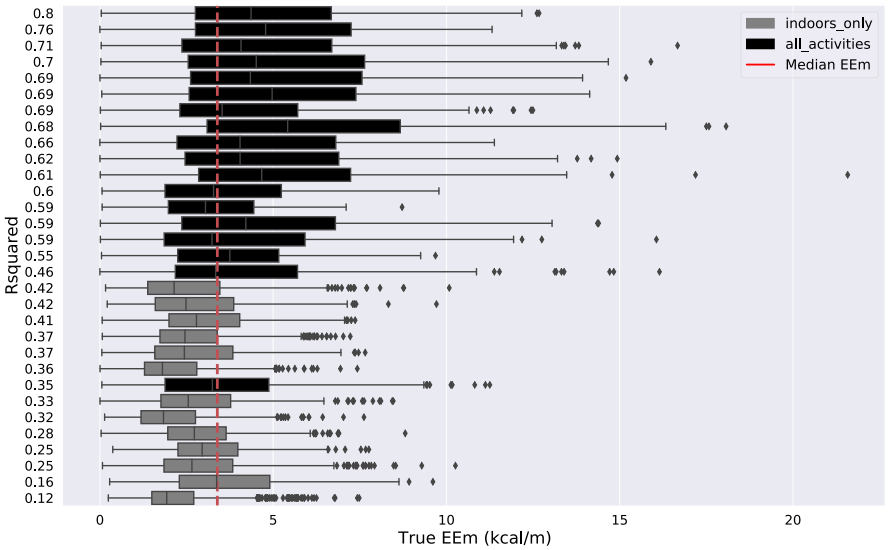

Fig. 10 Box plots of mean True EEm per participant ordered by R^2

with longer range of EEm values, the model does capture the variation. This may be due to the fact that for low EEm values, even a small RMSE error can lead to lower R^2. We indeed observe lower RMSE for indoor activities, while they produce lower R^2 values. We conclude that the proposed RNN model on average performs reasonably well for both high and low intensity activities (see Fig. 8). Because estimating PAEE from sedentary or low-intensity activities is considered challenging (van Hees et al. 2009), our proposed RNN model makes an important contribution.

6 Conclusion and future work

In this paper, we developed and tested a recurrent neural network architecture based on an efficient down-sampling method that incorporates standard deviation for down-sampling the input data to estimate physical activity energy expenditure within an elderly population. This approach is based on accelerometers at two body locations (wrist and ankle) and is able to take advantage of long time windows of predictor data (2 min to predict reasonably accurately the PAEE of older individuals. Moreover, the inclusion of participant-level data like age, gender and body composition further improved the accuracy of PAEE estimation, especially in activities with lower intensity ranges.

In summary, the results of this study demonstrate that RNNs incorporating GRU layers can solve the challenge of PAEE estimation. While they do not require any complex feature construction steps and can be trained with lower-resolution accelerometer data, if this is down-sampled with statistical dispersion metrics, RNNs produce PAEE estimations similar or better than competing methods. Because RNNs take into account longer windows in activity history without increasing the size and dimensionality of

their input, we believe that such modeling techniques are attractive when applied to free-living accelerometer data that is collected in a continuous way. Subsequently, our proposed down-sampling using statistical dispersion metrics (like standard deviation) proves to be really efficient since we achieved better results using ten times less data compared to averaging. Additionally, this strategy gives us the advantage of incorporating longer windows of prior sensor data with lower computational cost which also lead to better PAEE estimation. Finally, adding participant-level data (age, weight, height) when training our model can improve PAEE estimation significantly.

During the development of our models, we realized that the GOTOV dataset involves some data collection limitations. Indirect calorimetry was collected in a continuous way with only small breaks in between (max 1 minute). The rather small breaks between activities might make it difficult to estimate the EEm outcome per specific activity due to the energy expenditure's lag effect. In detail, without long discriminating breaks between activities, it is likely that past activities influence the EEm records of future ones. Additionally, we did not randomise the order of activities which might have introduced a slight bias in our training. For these reasons, it would have been interesting to test our findings in other similar labeled datasets with older individuals. However, to the best of our knowledge, GOTOV is as yet the only publicly available PAEE dataset with a focus on older individuals. Summarising, if there is no need to predict PAEE per specific activity, such as in our setting, this data collection can be used nicely to represent free-living conditions.

The RNN modeling advantage enabled taking into account preceding activity information by incorporating data of longer windows and letting the model decide on which information to emphasize on. The great advantage of the GOTOV dataset is that there are a satisfactory number of participants and that this data set is dedicated to people over 60 years of age. Because PAEE monitoring within the elderly might help stimulate vital and healthy ageing, the GOTOV dataset is perfectly suited for the development of activity recognition and PAEE estimation models.

Applying such a model to free-living data collections was one of the motivations of our study. In our future work, we intent to apply our modeling technique to physical activity and lifestyle improvement intervention studies on older individuals. From such an application, we envision that better insights in energy expenditure of older people will contribute to better physical behaviour guidelines for them to stay healthy, and potentially further stimulate vital and healthy ageing. In order to achieve this, we aim to build characteristic features of PAEE levels and PA types of long time periods (weeks, months) and relate them with parameters of metabolic health, general health and well-being. These relations between life style and health can then be turned into distinct recommendations for effectively maintaining mobility among older adults and a continuous monitoring system to track the adherence and improvement of metabolic health.

A Table of Abbreviations across the paper

Table 14 Abbreviations across the paper in alphabetical order

Abbreviations	
a	GENEActiv ankle
ANN	Artificial neural network
BMI	Body mass index
CNN	Convolutional neural network
CNN_{2min}	CNN trained with 2 min window of data
CNN_{5sec}	CNN trained with 5 s window of data
DLW	Doubly labeled water technique
EE	Energy expenditure
EEh	Energy expenditure per hour
EEm	Energy expenditure per minute
GOTOV	Growing old together Validation study
GRU	Gated recurrent unit
GRU_{AC}	GRU trained with accelerometer and activity classes data
GRU_{ID}	GRU trained with accelerometer and participant-level data
GRU_{ID_AC}	GRU with accelerometer, participant-level and activity classes data
HR	Heart rate
inR^2	Indoors R-squared
inRMSE	Indoors root mean squared error
IQR	Interquartile range
K4	COSMED $K4b^2$
LM	Linear model
LMM	Linear mixed model
LOSO-CV	Leave one subject out cross validation
LSTM	Long short-term memory
LUMC	Leiden University medical center
MET	Metabolic equivalent of task
MSE	Mean squared error
$outR^2$	Outdoors R-squared

Table 14 continued

Abbreviations	
outRMSE	Outdoors root mean squared error
PA	Physical activity
PAEE	Physical activity energy expenditure
PD	Percentile difference
REE	Resting energy expenditure
ReLU	Rectified linear unit
RF	Random forest
RFr	Random forest regressor
RMSE	Root mean squared error
RNN	Recurrent neural network
SD	Standard deviation
SR	Sampling rate
SVM	Signal vector magnitude
t	Average training time
TEE	Total energy expenditure
TEF	Thermic effect of food
VCO_2	Volume of carbon dioxide
VO_2	Volume of oxygen
w	GENEActiv wrist

B Open-source code

To make our experiments, scripts and results accessible to other researchers (open-source) we share them on the following git repository, https://github.com/parastelios/GOTOV_PAEE_publication.

References

Altini M, Penders J, Vullers R, Amft O (2015) Estimating energy expenditure using body-worn accelerometers: a comparison of methods, sensors number and positioning. IEEE J Biomed Health Inform 19(1):219–26. https://doi.org/10.1109/jbhi.2014.2313039

Bonomi AG, Plasqui G, Goris A, Westerterp KR (2009) Improving assessment of daily energy expenditure by identifying types of physical activity with a single accelerometer. J Appl Physiol 107(3):655–661. https://doi.org/10.1152/japplphysiol.00150.2009

Caron N, Peyrot N, Caderby T, Verkindt C, Dalleau G (2020) Estimating energy expenditure from accelerometer data in healthy adults and patients with type 2 diabetes. Exp Gerontol 134:110894. https://doi.org/10.1016/j.exger.2020.110894

Chen L, Fox R, Ku P, Sun W, Chou P (2012) Prospective associations between household-, work-, and leisure-based physical activity and all-cause mortality among older taiwanese adults. Asia Pac J Public Health 24(5):795–805

Chung J, Gülçehre Ç, Cho K, Bengio Y (2015) Gated feedback recurrent neural networks. In: Proceedings of the 32nd international conference on international conference on machine learning (ICML), Lille, France, 6–11 July 2015

Cicero AF, D'Addato S, Santi F, Ferroni A, Borghi C, Brisighella HS (2012) Leisure-time physical activity and cardiovascular disease mortality: the Brisighella heart study. J Cardiovasc Med (Hagerstown) 13(9):559–64. https://doi.org/10.2459/JCM.0b013e3283516798

Dong B, Biswas S, Montoye A, Pfeiffer K (2013) Comparing metabolic energy expenditure estimation using wearable multi-sensor network and single accelerometer. In: 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC). https://doi.org/10.1109/EMBC.2013.6610138

Edel M, Köppe E (2016) Binarized-blstm-rnn based human activity recognition. In: 2016 international conference on indoor positioning and indoor navigation (IPIN), pp 1–7. https://doi.org/10.1109/IPIN.2016.7743581

Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S (2014) A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol Meas 35(11):2191–2203

Frisard MI, Broussard A, Davies SS, Roberts LJn, Rood J, de Jonge L, Fang X, Jazwinski SM, Deutsch WA, Ravussin E (2007) Louisiana healthy aging study aging, resting metabolic rate, and oxidative damage: results from the Louisiana healthy aging study. J Gerontol: Ser A 62(7):752–9. https://doi.org/10.1093/gerona/62.7.752

Gjoreski H, Kaluža B, Gams M, Milić R, Luštrek M (2013) Ensembles of multiple sensors for human energy expenditure estimation. In: Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing, UbiComp '13, New York, NY, USA. Association for Computing Machinery

Guan Y, Plötz T (2017) Ensembles of deep lstm learners for activity recognition using wearables. Proc ACM Interact, Mob, Wearab Ubiquitous Technol 1(2):1–28. https://doi.org/10.1145/3090076

Hills AP, Mokhtar N, Byrne NM (2014) Assessment of physical activity and energy expenditure: an overview of objective measures. Front Nutr 1:5. https://doi.org/10.3389/fnut.2014.00005

Hochreiter S (1998) The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertain Fuzziness Knowl-Based Syst 06(02):107–116. https://doi.org/10.1142/S0218488598000094

Hortobágyi T, Mizelle C, Beam S, DeVita P (2003) Old adults perform activities of daily living near their maximal capabilities. J Gerontol A Biol Sci Med Sci 58(5):453–60. https://doi.org/10.1093/gerona/58.5.m453

Jones L, Waters D, Legge M (2009) Walking speed at self-selected exercise pace is lower but energy cost higher in older versus younger women. J Phys Act Health 6(3):327–32

Keys A, Taylor HL, Grande F (1973) Basal metabolism and age of adult man. Metabolism 22(4):579–87. https://doi.org/10.1016/0026-0495(73)90071-1

Kim ZM, Oh H, Kim H-G, Lim C-G, Oh K-J, Choi H-J (2017) Modeling long-term human activeness using recurrent neural networks for biometric data. BMC Med Inform Decis Mak 17:57

Kingma DP, Ba J (2015)Adam: a method for stochastic optimization. In: 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, Conference Track Proceedings

Knaggs JD, Larkin KA, Manini TM (2011) Metabolic cost of daily activities and effect of mobility impairment in older adults. J Am Geriatr Soc 59(11):2118–23. https://doi.org/10.1111/j.1532-5415.2011.03655.x

Lee K, Park C, Kim N, Lee J (2018) Accelerating recurrent neural network language model based online speech recognition system. In: 2018 IEEE international conference on acoustics, speech and signal processing, ICASSP 2018, Calgary, AB, Canada, April 15–20, 2018

Leonard W (2012) Laboratory and field methods for measuring human energy expenditure. Am J Hum Biol 24(3):372–84. https://doi.org/10.1002/ajhb.22260

Li S, Xu J (2018) A recurrent neural network language model based on word embedding. In: Web and big data—APWeb-WAIM 2018 international workshops: MWDA, BAH, KGMA, DMMOOC, DS, Macau, China, July 23–25, 2018, Revised Selected Papers

Liu S, Gao RX, Freedson PS (2012) Computational methods for estimating energy expenditure in human physical activities. Med Sci Sports Exerc 44(11):2138–46

Lyden K, Kozey SL, Staudenmeyer JW, Freedson PS (2011) A comprehensive evaluation of commonly used accelerometer energy expenditure and met prediction equations. Eur J Appl Physiol 111(2):187–201

Manini TM, Everhart JE, Patel KV, Schoeller DA, Colbert LH, Visser M, Tylavsky F, Bauer DC, Goodpaster BH, Harris TB (2006) Daily activity energy expenditure and mortality among older adults. J Am Med Assoc (JAMA) 296(2):171–9. https://doi.org/10.1001/jama.296.2.171

Mardini MT, Nerella S, Wanigatunga AA, Saldana S, Casanova R, Manini TM (2020) Deep chores: estimating hallmark measures of physical activity using deep learning. In: AMIA annual symposium proceedings. AMIA Symposium, pp 803–812

Martin KR, Koster A, Murphy RA, Van Domelen DR, Hung M-y, Brychta RJ, Chen KY, Harris TB (2014) Changes in daily activity patterns with age in U.S. men and women: national health and nutrition examination survey 2003–04 and 2005–06. J Am Geriatr Soc 62(7):1263–71. https://doi.org/10.1111/jgs.12893

McLaughlin JE, King GA, Howley ET, Bassett DR, Ainsworth BE (2001) Validation of the cosmed k4 b2 portable metabolic system. Int J Sports Med 22(4):280–4. https://doi.org/10.1055/s-2001-13816

Montoye A, Begum M, Henning Z, Pfeiffer KA (2017) Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. Physiol Meas 38(2):343–57. https://doi.org/10.1088/1361-6579/38/2/343

Montoye A, Conger SA, Connolly CP, Imboden MT, Nelson MB, Bock JM, Kaminsky LA (2017) Validation of accelerometer-based energy expenditure prediction models in structured and simulated free-living settings. Meas Phys Educ Exerc Sci 21(4):223–234. https://doi.org/10.1080/1091367X.2017.1337638

Okai J, Paraschiakos S, Beekman M, Knobbe A, de Sá CR (2019) Building robust models for human activity recognition from raw accelerometers data using gated recurrent units and long short term memory neural networks. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). https://doi.org/10.1109/EMBC.2019.8857288

O'Driscoll R, Turicchi J, Hopkins M, Horgan GW, Finlayson G, Stubbs JR (2020) Improving energy expenditure estimates from wearable devices: a machine learning approach. J Sports Sci. https://doi.org/10.1080/02640414.2020.1746088

Paraschiakos S, Cachucho R, Moed M, van Heemst D, Mooijaart S, Slagboom EP, Knobbe A, Beekman M (2020) Activity recognition using wearable sensors for tracking the elderly. User Model User-Adap Int 30:567–605. https://doi.org/10.1007/s11257-020-09268-2

Petersen CB, Gronbaek M, Helge JW, Thygesen LC, Schnohr P, Tolstrup JS (2012) Changes in physical activity in leisure time and the risk of myocardial infarction, ischemic heart disease, and all-cause mortality. Eur J Epidemiol 27(2):91–9. https://doi.org/10.1007/s10654-012-9656-z

Roberts SB, Dallal GE (2005) Energy requirements and aging. Public Health Nutr 8(7a):1028–1036. https://doi.org/10.1079/PHN2005794

Staudenmayer J, Pobe rD, Crouter S, Bassett D, Freedson P (2009) An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J Appl Physiol 107(3):1300–7. https://doi.org/10.1152/japplphysiol.00465.2009

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R (2015) Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12(3):1–10. https://doi.org/10.1371/journal.pmed.1001779

Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, Mottram A, Meyer C, Ravuri S, Protsyuk I, Connell A, Hughes CO, Karthikesalingam A, Cornebise J, Montgomery H, Rees G, Laing C, Baker CR, Peterson K, Reeves R, Hassabis D, King D, Suleyman M, Back T, Nielson C, Ledsam JR, Mohamed S (2019) A clinically applicable approach to continuous prediction of future acute kidney injury. Nature 572:116–119. https://doi.org/10.1038/s41586-019-1390-1

van Ballegooijen AJ, van der Ploeg HP, Visser M (2019) Daily sedentary time and physical activity as assessed by accelerometry and their correlates in older adults. Eur Rev Aging Phys Act. https://doi.org/10.1186/s11556-019-0210-9

van de Rest O, Schutte BAM, Deelen J, Stassen SAM, van den Akker EB, van Heemst D, Dibbets-Schneider P, van Dipten-van der Veen RA, Kelderman M, Hankemeier T, Mooijaart SP, van der Grond J, Houwing-Duistermaat JJ, Beekman M, Feskens EJM, Slagboom PE (2016) Metabolic effects of a 13-weeks lifestyle intervention in older adults: the growing old together study. Aging 8(1):111–124. https://doi.org/10.18632/aging.100877

van Hees VT, van Lummel RC, Westerterp KR (2009) Estimating activity-related energy expenditure under sedentary conditions using a tri-axial seismic accelerometer. Obesity 17(6):1287–1292. https://doi.org/10.1038/oby.2009.55

Volchan SB (2002) What is a random sequence? Am Math Mon 109(1):46–63. https://doi.org/10.1080/00029890.2002.11919838

Weinsier RL, Schutz Y, Bracco D (1992) Reexamination of the relationship of resting metabolic rate to fat-free mass and to the metabolically active components of fat-free mass in humans. Am J Clin Nutr 55(4):790–4. https://doi.org/10.1093/ajcn/55.4.790

Weir JDV (1949) New methods for calculating metabolic rate with special reference to protein metabolism. J Physiol 109(1–2):1–9. https://doi.org/10.1113/jphysiol.1949.sp004363

Westendorp RG, van Heemst D, Rozing MP, Frölich M, Mooijaart SP, Blauw GJ, Beekman M, Heijmans BT, de Craen AJ, Slagboom PE, Leiden Longevity Study Group (2009) Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: the Leiden longevity study. J Am Geriatr Soc 57(9):1634–37.https://doi.org/10.1111/j.1532-5415.2009.02381.x

Wijsman CA, Westendorp RG, Verhagen EA, Catt M, Slagboom PE, de Craen AJ, Broekhuizen K, van Mechelen W, van Heemst D, van der Ouderaa F, Mooijaart SP (2013) Effects of a web-based intervention on physical activity and metabolism in older adults: randomized controlled trial. J Med Internet Res 15(11):e233. https://doi.org/10.2196/jmir.2843

Zhu J, Pande A, Mohapatra P, Han JJ (2015) Using deep learning for energy expenditure estimation with wearable sensors. In 2015 17th international conference on E-health networking, application services (HealthCom). https://doi.org/10.1109/HealthCom.2015.7454554