Activity recognition using wearable sensors for tracking the elderly

8 authors, including:

Stylianos Paraschiakos
Leiden University Medical Centre
4 PUBLICATIONS **1** CITATION

SEE PROFILE

Ricardo Cachucho
Leiden University
16 PUBLICATIONS **89** CITATIONS

SEE PROFILE

Diana van Heemst
Leiden University Medical Centre
367 PUBLICATIONS **8,704** CITATIONS

SEE PROFILE

Simon P Mooijaart
Leiden University Medical Centre
254 PUBLICATIONS **5,266** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Care for older people with (end stage) kidney disease View project

GEHA linkage analyses for human longevity View project

Activity recognition using wearable sensors for tracking the elderly

Stylianos Paraschiakos, et al. *[full author details at the end of the article]*

© The Author(s) 2020

Abstract

A population group that is often overlooked in the recent revolution of self-tracking is the group of older people. This growing proportion of the general population is often faced with increasing health issues and discomfort. In order to come up with lifestyle advice towards the elderly, we need the ability to quantify their lifestyle, before and after an intervention. This research focuses on the task of activity recognition (AR) from accelerometer data. With that aim, we collect a substantial labelled dataset of older individuals wearing multiple devices simultaneously and performing a strict protocol of 16 activities (the GOTOV dataset, $N = 28$). Using this dataset, we trained Random Forest AR models, under varying sensor set-ups and levels of activity description granularity. The model that combines ankle and wrist accelerometers (GENEActiv) produced the best results (accuracy > 80%) for 16-class classification. At the same time, when additional physiological information is used, the accuracy increased (> 85%). To further investigate the role of granularity in our predictions, we developed the LARA algorithm, which uses a hierarchical ontology that captures prior biological knowledge to increase or decrease the level of activity granularity (merge classes). As a result, a 12-class model in which the different paces of walking were merged showed a performance above 93%. Testing this 12-class model in labelled free-living pilot data, the mean balanced accuracy appeared to be reasonably high, while using the LARA algorithm, we show that a 7-class model (lying down, sitting, standing, household, walking, cycling, jumping) was optimal for accuracy and granularity. Finally, we demonstrate the use of the latter model in unlabelled free-living data from a larger lifestyle intervention study. In this paper, we make the validation data as well as the derived prediction models available to the community.

Published online: 23 June 2020

Springer

1 Introduction

With the recent availability of a wide range of affordable personal sensing devices, a growing group of people is starting to experiment with *self-tracking*. This new development, aimed at quantifying various aspects of one's personal life, is allowing people to better understand their life, improve their professional efficiency, or optimise a range of health factors. While it has become a lot easier to gather large quantities of raw data about one's daily activities and physiological parameters, making sense of all this data is still a considerable challenge. Especially when better health is the primary goal, understanding the exact nature of one's activities, in other words, placing the raw measurements into context, is a crucial prerequisite for further processing of the data.

Although the self-tracking movement has become a mature field of study, the primary subject of study is still the healthy, young individual that is interested in how technology can contribute to a better life. In terms of available models and data, there is a clear focus on younger individuals, which produces a bias in the physiological nature of the data, as well as the typical lifestyle recorded. A systematic approach towards self-tracking aimed at the elderly has so far been under-represented (Schrack et al. 2016), which is unfortunate since, in most countries, the elderly are becoming a larger proportion of the population.

Among senior citizens, the biggest potential of self-tracking lies in improving or maintaining their general health. A major determinant in that is one's lifestyle, so quantifying that is the first priority. In fact, many medical studies of healthy ageing are aimed at lifestyle changes (Schrack et al. 2016; van de Rest et al. 2016; Jansen 2015; Wijsman et al. 2013). Typically, senior participants are encouraged to replace a mostly sedentary lifestyle by a more active one, for example by proposing a daily moderate exercise. As part of these lifestyle-oriented intervention studies, a pre- and post-intervention period of free-living is recorded, in order to quantify the change in the participant's activity profile. We have been involved in several such studies in the Netherlands, and a recurring theme in these studies was the objective measurement of activity type and level by means of wearable sensors.

To investigate such conditions in older people, in the Netherlands, several observational and experimental biomedical studies were initiated, with a focus on *healthy ageing*. These studies aimed to identify both genetic and environmental determinants of healthy ageing and included the collection of various types of physical activity data. As an example, the GOTO study (van de Rest et al. 2016) (one of the Leiden biomedical ageing studies that follow people predisposed to become long-lived and age healthily) studied 164 older participants (mean age 63.2 years). Alongside several clinical parameters, gene expression levels and MRI scans, the participant's daily activities using wearable sensors were measured over 2 weeks, one before and one after the lifestyle intervention. A major hurdle in the integrated analysis of this data is the correct interpretation of the sensor data since it is known in considerable detail what accelerations different body parts were experiencing, but *what exactly the participants were doing* is not

immediately apparent. Similar challenges exist in other studies, including some of our studies, which we list in "Appendix 2" (Westendorp et al. 2009; van de Rest et al. 2016; Jansen 2015; Wijsman et al. 2013).

In the absence of good training data and predictive models aimed at the activity recognition of the elderly using wearables, we decided to conduct our own validation study using a wide range of sensors and building appropriate models. In this paper, we report on the *Growing Old TOgether Validation* (GOTOV) study. We collected labelled wearable sensor data of 35 subjects, aged between 60 and 85 with a mean age 65 and mean BMI 27 kg/m^2. These population details are similar to the above-mentioned studies performed in Leiden University Medical Center (LUMC), like the GOTO, where different participants took place with a mean age 63.2 and mean BMI 27 kg/m^2. This dataset encompasses 16 activities with low (lying down, sitting), mid (standing, household activities) and high (walking and cycling) levels of intensity. The data collection protocol was designed to include activities that cover all periods of everyday living for this age group in approximately 90 min.

The GOTOV data collection protocol was developed to mostly recreate free-living conditions. This means that in order to be representative, the data collection took place in two separate moments. First, low and mid-intensity activities were measured indoors in a laboratory environment. This part of the protocol simulates activities such as sitting on a sofa while reading a newspaper or performing different household activities. The second part of the protocol was measured outdoors, where we collected data of more complex and high-intensity activities like walking and cycling during normal urban conditions. Adding to that, some activities can be specified at different granularity.

The GOTOV study was designed with two goals in mind, (1) to encompass all the sensor set-ups involved in four biomedical studies in Leiden ("Appendix 2"), and (2) to involve novel sensors for capturing additional physiological parameters (e.g. heart and breath rate variability, skin temperature, etc.), which could then be correlated with the more comfortable and affordable wearables such as accelerometers. Since individual studies required different subsets of the data, separate models had to be trained for different subsets of the data. For example, the GOTO study involves a moderate set-up of only accelerometers worn at the wrist and the ankle, while for example Switchbox (Jansen 2015) including heart and breath rate sensors. The interplay between sensors and their body location, the granularity of activity classes and resulting model accuracies forms a large part of the experimental results that we present here.

In our study, we introduce an activity ontology, formulated during the design of the data collection protocol, which consists of three levels of granularity. Having these levels allows one to recognise the activity at several levels of specificity. In fact, there is a trade-off between the granularity and accuracy of the model: as you move down the hierarchy, making detailed distinctions between sub-classes requires better data and more accurate models, a luxury that is not always available. We propose a new learning method (LARA, for *Learning activity recognition models accurately*) that aims to guarantee a certain, user-specifiable, level of accuracy, by moving up the hierarchy (from specific to general) until the desired accuracy is achieved.
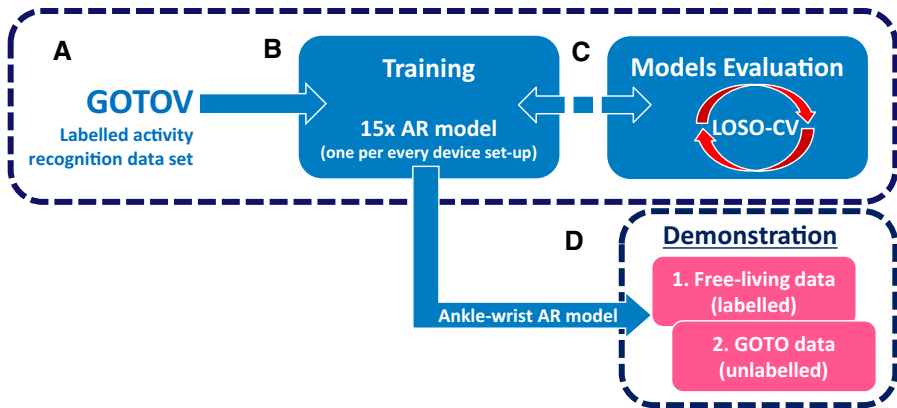
Fig. 1 Study Flow-chart

At each iteration, the granularity of the most problematic part of the hierarchy is reduced.

The method proposed here is agnostic to the actual activity recognition (AR) algorithm employed, so any prediction algorithm can be plugged in. But for the experiments presented below, we are using a novel combination of two algorithms that are able to deal with the specific challenges of AR data, Accordion algorithm and random forest (RF). The Accordion (Cachucho et al. 2014) algorithm specifically extracts features from temporal data that are suitable for further classification of the activities by a classification algorithm, RF here. Since raw sensor data are captured at a relatively high frequency, and activities are only predicted, say, once per second, the sensor data will need to be aggregated over windows of various lengths (to capture both short and long-term effects). Having extracted promising features with Accordion, we employ a random forest (Breiman 2001) to produce the actual prediction.

Summarising, the main contributions of this paper are as follows:

1. We present and make publicly available a new data collection focused on the elderly which incorporates 16 activities following a three-level hierarchical ontology.
2. We develop a pipeline combining Accordion and RF with a novel method, called LARA, that systematically considers different levels of granularity appropriate for the sensor set-up in question.
3. We deliver robust and highly accurate activity recognition models.
4. We report on optimal sensor set-ups in terms of size, sensor type and body location.
5. We demonstrate the performance of our models in both labelled and unlabelled free-living accelerometer data.

In Fig. 1, we present our study workflow. This is divided into 4 parts, i.e. A, B, C, D. (A) represents the labelled AR dataset, that will become publicly available

(GOTOV), (B) the pipeline of the consecutive AR tools (Accordion, Random Forest, LARA), (C) the evaluation of the AR models using the Leave-One-Subject-Out cross-fold validation (LOSO-CV), and (D) the demonstration of one of our models, with free-living data from two sources: 1. a labelled collection and 2. the GOTO lifestyle intervention study (unlabelled).

In detail the paper's structure, in Background (Sect. 2), we present the related work and put our contributions into perspective. In LARA (Sect. 3), we introduce some mathematical notation followed by the learning method LARA that we propose. In Sect. 4, we introduce part of **A** (Fig. 1) including the data collection protocol, the activities ontology and a description of the dataset made publicly available. This is followed by part **B** and **C** of the study workflow (Fig. 1), in Sect. 5 (training of AR models) where we present the experiments, discuss the feature construction step, optimise the classification model and present the optimal location of the sensors. At the end of the section, we also evaluate and demonstrate the use of one of our models in labelled free-living data (part **D-1**, Fig. 1) and in free-living data of the GOTO study (part **D-2**, Fig. 1). In Sect. 6, we present our conclusions and future work. Finally, in the Appendix we add a list of all the abbreviations ("Appendix 1") and the details of the four Leiden biomedical ageing studies ("Appendix 2").

2 Background

In recent years, the advances in wearable sensor technology (smaller size, non-invasive, long battery life) and its potential application to studies that monitor physical activity have attracted wide attention. Since physical activity is one of the main determinants of healthy ageing, from early 2000 until today, there is a significant increase in ageing studies using such devices (Shiroma et al. 2018). The (Murphy 2009; Schrack et al. 2016; Tedesco et al. 2017; Shiroma et al. 2018) review papers give an extensive comparison of up-to-date studies using wearables or/and stationary sensors to monitor physical activity in older adults. In this section, considering wearables sensors as predictors, we will have a systematic analysis that puts our study into perspective with others, with special attention to activity recognition in older people.

Comparing the literature, it can be seen that identifying and quantifying activity (moderate to vigorous intensity) of older adults from sensor data is a challenging problem. Most studies tend to use activity metrics such as cut points and/or metabolic equivalents (METs) thresholds (Rejeski et al. 2016; Schrack et al. 2018). However, most of these threshold values are calculated in younger individuals and may prove problematic when applied to older ones since activities patterns differ depending on the age (Sallis 2000; Martin et al. 1992, 2014). Additionally, if they are not accompanied by a daily activity log, they are more or less disconnected from the activity's context. As a result, it is hard to identify which activities are linked with health parameters and may, therefore, be used as healthy ageing recommendations. For this reason, we would like to use AR models to investigate if there is a relationship between the different activities and health parameters.

Activity recognition (AR) using wearable sensors (e.g. accelerometers) could be a solution to this challenge. In the last 20 years already, the field of AR has significantly evolved, with research both in using wearable sensors (Bao and Intille 2004; Parkka et al. 2006; Patel et al. 2009; Altun and Barshan 2010; Lara and Labrador 2013) and environment-based sensors (Ranasinghe et al. 2016; Rashidi and Mihailidis 2013). However, the majority of the relevant literature (Bao and Intille 2004; Parkka et al. 2006; Altun and Barshan 2010; Reiss and Stricker 2012; Attal et al. 2015; Saez et al. 2016) focuses on developing AR models mainly on younger populations (mid 20s). Adding to that, the few studies focused on older individuals they either study participants with various health conditions (Patel et al. 2009; Gao et al. 2014) or when they focus to healthy older individuals they use stationary sensors (Moshtaghi et al. 2015; Duong et al. 2005). This concludes in making the models non-generalisable to wider healthy older population. In more detail, in Del Rosari et al. (2014) it is presented that training AR models in a young population and applied to an older one (and vice versa) conclude to lower performance compared to models trained and tested on the same population. As a result, existing AR models which were developed on a younger population could not be generalisable to an older one. For that reason, it is important to study physical activity and develop AR models focused on healthy older populations.

Additionally, enhancing reproducibility in the field of AR is a major challenge (Brush et al. 2011). The majority of the publications in AR are not reproducible, mainly because both algorithms and dataset are not publicly available. Take as an example the publications on activity recognition for older people: none of them has made the datasets publicly available; even not the dataset of the most cited AR publication (Bao and Intille 2004). In contrast, there are few publications (Altun and Barshan 2010; Reiss and Stricker 2012; Chavarriaga et al. 2013) that have published their collected datasets or parts of them. However, these publicly available datasets fall short to support studies on ageing populations. Therefore, we introduce a new publicly available activity recognition dataset that focuses on older individuals (between 60 and 85 with a mean age 65). The collected dataset consists of 35 participants, 2 different devices (GeneActivs and Equivital) on 3 different body locations (ankle, wrist, chest) and 16 activity classes, with a hierarchical ontology.

From a modelling perspective, focusing on wearables, due to the fact that different activities utilise different body parts, it has been proven that AR model's performance is highly dependent on the sensors' body position (Bao and Intille 2004). Moreover, different studies train AR models using multiple wearables at different body locations (chest, wrist, ankle, thigh, etc.). Consequently, there is a need for studies that could recognise daily activities independently of the body position (Khan et al. 2010). Due to our hierarchical activity ontology design, we create activity detail specific models per sensor combination. This way, depending on the desired activity detail prediction, we can suggest a sensor combination.

The use of an ontology structure for activity recognition has been introduced before in several studies, both for vision-based methods and for wearables sensors-based, see details in this review (Chen and Nugent 2009). More specific, ontologies have been used to represent the hierarchy of different activities (Riboni and Bettini 2011), which allows to describe the activities in a more structured way and identify

activities with different readings because of movement, body posture or intensity as one.

Adding to that, we have seen several studies (Reiss and Stricker 2012; Parkka et al. 2006; Dutta et al. 2016) that started with a data collection protocol that includes a big number of activities, but in the end, the reported models are tested only on a subset of them. Often, these activity classes are merged into a higher concept activity-class based on confusion (Dutta et al. 2016). However, this step is not clearly explained or justified. With the use of a prior ontology which is used by an algorithm as an input (LARA), this merging of classes is performed and achieved in a more methodological way.

Adding to body locations and ontology, there is a lack of standardised activity protocols that could generalise laboratory-trained AR models to free-living collected data. Sasaki et al. (2016) compares classification algorithms for accelerometer data in the laboratory and in free-living conditions. The results prove that laboratory-trained AR models will underperform when they are used to classify activity from free-living data. On the other hand, creating free-living labelled datasets can be challenging, since participants should record daily logs or use extra devices, like cameras, increasing the studies cost and interfering with the participant's private life. Therefore, there is a need for methods that develop accurate AR models, that are robust under free-living conditions.

3 LARA: learning activity recognition models accurately

In this section, we introduce fundamental concepts to understand the created models and datasets that are used to create such models.

3.1 Preliminaries

We will start by focusing on data produced by a collection of sensors. Such a collection produces datasets that can be formalised as a multivariate time series matrix. We define a multivariate time series as an $n \times m$ matrix T (see Eq. 1).

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_{1,1} & \mathbf{T}_{1,2} & \cdots & \mathbf{T}_{1,m} \\ \mathbf{T}_{2,1} & \mathbf{T}_{2,2} & \cdots & \mathbf{T}_{2,m} \\ \vdots & \vdots & \mathbf{T}_{i,j} & \vdots \\ \mathbf{T}_{n-1,1} & \mathbf{T}_{n-1,2} & \cdots & \mathbf{T}_{n-1,m} \\ \mathbf{T}_{n,1} & \mathbf{T}_{n,2} & \cdots & \mathbf{T}_{n,m} \end{pmatrix} \tag{1}$$

Note that this matrix, is on a temporal order: $\mathbf{T}_{i,j}$ represents a measurement at time point i for variable j, e.g. the input value of X-axes of an accelerometer at the time-point i. This collection can be decomposed into subsets of sensors $s \in \mathcal{S}$, where $\mathcal{S}$ is the set of all sensors employed. A subset of sensors $s \in \mathcal{S}$ refers to one or more sensors combined, which represent a sub-matrix of the predictors matrix, $\mathbf{T}_s \subseteq \mathbf{T}$. We refer to $\mathbf{T}_s$ as sensor set-up.

With $\mathbf{T}$ being a time series, the indexes $\{1, \ldots, n\}$ of $\mathbf{T}$ represents a temporal order and each point of a time series is collected at a specific moment in time. We refer to each of these moments as *timestamps* $\omega(i)$. From the timestamps, one can derive relevant information such as sampling rates. The sampling rate of $\mathbf{T}$ can be calculated as follows:

$$r_{\mathbf{T}} = \frac{n-1}{\omega(n) - \omega(1)}. \tag{2}$$

Assume that associated with the multivariate time series $\mathbf{T}$, we have a categorical dependent variable $\mathbf{c}$ representing the target variable (subject's activity). In the case of time series, this target variable is a vector $\mathbf{c} = (c_1, \ldots, c_{|\mathbf{c}|})$, with temporal order. The target vector $\mathbf{c}$ represents a finite sequence of classes, with the following properties:

- Target variable $\mathbf{c}$ has fewer or the same number of instances than predictors matrix $\mathbf{T}$: $|\mathbf{c}| \leq n$.
- $\mathbf{c}$ and $\mathbf{T}$ are collected during the same period of time: $\omega(|\mathbf{c}|) = \omega(n)$.
- $c_{i'}$ denotes a class by means of an integer $\{1, \ldots, k\}$, where k is the number of classes;

Please note that the categorical target vector $\mathbf{c}$ has a lower sampling rate than $\mathbf{T}$, $r_{\mathbf{c}} \leq r_{\mathbf{T}}$.

3.2 Activity recognition problem

In our activity recognition framework, $\mathbf{c}$ is a target vector where the classes represent different activities. k represents the number of human activities ($k = 16$). Given a set of activities $A = \{a_1, \ldots, a_k\}$, we define an *ontology* as a sequence of partitions $P_0, \ldots P_d$, where the parameter d refers to the *depth* of the ontology (the number of levels minus one) and

- for $d = 0$: P_0 is the singleton consisting of A: $P_0 = \{A\}$,
- for $d = l$: P_l aggregates a subset of activities P_{l+1} in a parent activity: a, where $\forall a \in P_l, \exists G \subset P_{l+1} : a = \bigcup G$.

The partitions P_l represent the *levels* of the ontology: different levels of activity description, ranging from individual activities (P_d) to all activities merged together (P_0). This definition of an ontology allows multiple activities with any number of levels of aggregation depending on their granularity.

In this paper, the specific ontology used is of depth $d = 3$, where levels P_1, P_2 and P_3 are presented in Fig. 2, with P_3 the classes of the bottom row of the ontology, P_2 the intermediate level (lying down, sitting, standing, household, jumping, walking, cycling), and P_1 the level representing activities at three levels of intensity (low, medium and high). In Sect. 4.1.2, we explain the reasoning behind our selected ontology.
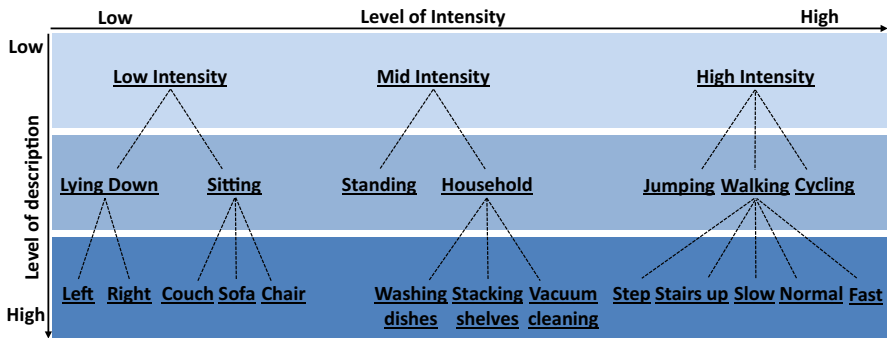
Fig. 2 Activities hierarchical ontology. (Color figure online)

3.2.1 Feature construction and selection

When facing multiple sampling rates as stated above, one important aspect is how to map multiple instances of the independent variables **T** to one instance of the dependent variable **c**. The most common solution for this mapping is to take multiple values of **T** and aggregate them to one value. This technique is widely known as sliding windows (Bifet 2009). Such an aggregation procedure will result in what we call *aggregate features*.

Note that the possibilities of combinations in term of window size *w* and a potentially large set of aggregation functions will lead to a fairly large search space. In this paper, we use Accordion (Cachucho et al. 2014), a published algorithm for feature construction and selection in mixed-sampling rates time series data. In detail, using a set of aggregation functions and given a maximum window (here 5 sec) and a prediction sampling rate, it selects the best representative features and window sizes, guided by the information gain measure.

3.2.2 Learning accurate and robust AR models

Now assume that we are given a set of aggregate features $\mathcal{F}$ with a cardinality of *m*. Such a set can be represented as a $|\mathbf{c}| \times m$ feature matrix, which we call **F**. Given **c** and **F**, the definition of an activity recognition model is:

$$\hat{c}_{i'} = f(\mathbf{F}_{i'}), \tag{3}$$

In Eq. 3, *f* is a classification model learned by a classification algorithm and $\hat{c}_{i'}$ is a class prediction, for instance, i' based on row $\mathbf{F}_{i'}$ of matrix **F**.

Assume that we have a sensor set-up, producing a multivariate time series $\mathbf{T}_s$ and a target of human activities, $c_{i'}$. The method that we propose takes $\mathbf{T}_s$ and $c_{i'}$ as inputs, and outputs a classification model *f*, as described in Algorithm 1. This algorithm begins by attempting to build a model for predicting all activities separately (multi-class classification with *k* classes). As long as this model does not reach a specified accuracy threshold θ (accuracy is computed by the count of

correct predictions), the algorithm iteratively moves up the activity ontology until the minimum desired activity description P_{min} is reached, for our application level P_2. We named this method Learning Activity Recognition models Accurately (LARA) and use it to determine the appropriate number of classes that should be used for classification.

ALGORITHM 1: LARA: Learning Activity Recognition Accurately

Input: sensor set-up $\mathbf{T}_s$, a target vector $\mathbf{c}$ with k classes in an ontology $P_0, \ldots, P_d$, an accuracy threshold θ and the desired minimum level P_{min}
Output: Activity Recognition model f
Build aggregate features: $\mathbf{F} \leftarrow Accordion(\mathbf{T}_s)$
$depth \leftarrow d$
$p \leftarrow P_d$
repeat
    $count \leftarrow 0$
    for *each test subject in* $\mathbf{F}_{i'}$ **do**
        use the remaining subjects to learn model: $f \leftarrow randomForest(\mathbf{F}_{i'})$
        predict values on test subject: $\hat{c}_{i'} \leftarrow f(\mathbf{F}_{i'})$
        $count \leftarrow count + |\{i'|\hat{c}_{i'} = c_{i'}\}|$
    end
    if $\frac{count}{|\mathbf{c}|} \leq \theta$ **then**
        for *each* $a \in P_{depth-1}$ **do**
            if *activities in p corresponding to a are not yet merged* **then**
                compute $IGC(a)$
            end
        end
        merge activities in p corresponding to a with highest IGC in $P_{depth-1}$
        if $p = P_{depth-1}$ **then**
            $depth \leftarrow depth - 1$
        end
    end
until $depth = P_{min} \lor \frac{count}{|\mathbf{c}|} \geq \theta$;
return f

At each iteration, LARA merges one set of activities (siblings) that belong to a single parent activity (e.g. {*Lying Down Left*}, {*Lying Down Right*} becomes {*Lying Down*}). In terms of the ontology $P_0, \ldots, P_d$, the iterative algorithm starts with the individual activities $P = P_d$ at depth $depth = d$, considers all parent activities in $P_{depth-1}$ and merges the sibling activities of the parent that causes the most 'confusion' (explained below). After all activities on level $depth$ have been merged (assuming the accuracy threshold θ has not yet been reached), the activity partition P now equals $P_{depth-1}$. Then $depth$ is decreased by one, and the process naturally proceeds until sufficient accuracy or minimum $depth$ is reached.

In order to decide at each level which is the next set of activities to be merged into their parent a, we compute the *intra-group confusion* IGC. For a given parent a and its constituting activities $a_1, \ldots, a_k$, the IGC computes how many cases of a_i are predicted (confused) as a_j, for all $i \neq j$. In other words, LARA focus on confusion between classes of the same parent.

As an example, given our ontology (Fig. 2). In the first iteration with $k = 16$ classes and $depth = 3$, if a model does not reach the accuracy threshold θ, the LARA algorithm will:

1. Compute the IGC per group of siblings, e.g. the ICG of *lyingDown*'s children, *sitting*'s children, etc.
2. Identify the group with the highest ICG.
3. Merge the siblings to one class, the parent, e.g. *walkingSlow*, *walkingNormal* and *walkingFast* into class *walking*.
4. Train the new model with the given classes and evaluate if it reached the accuracy threshold θ.
5. If the threshold is not reached, it continues merging sibling classes per level until either the threshold or the minimum activity description is reached.

As part of the threshold criterion in Algorithm 1, the accuracy is computed using Leave-One-Subject-Out cross-validation (LOSO-CV) (Bao and Intille 2004; Patel et al. 2009; Reiss and Stricker 2012; Saez et al. 2016). This evaluation methodology splits the dataset into N folds, where N is the number of human subjects, using $N - 1$ folds as the training set and the remaining fold as the test set. This simulates how the models will be used in the future: on subjects not yet known to the model. Through this procedure, we guarantee that the classifier learns nothing about the subject to be predicted, thus decreasing learning bias and achieving reliable results for cross-subject prediction. Note that an evaluation methodology that does not use LOSO-CV would provide a more optimistic estimate of the accuracy. More specifically, with normal cross-validation, data from the same subject would be used in both training and test set leading to an overestimation in papers that report accuracies using it (Attal et al. 2015; Gao et al. 2014; Chavarriaga et al. 2013). Unlike cross-validation, this procedure estimates the expected accuracy, not only on new data but also on new subjects. The process is repeated for all the subjects and the average accuracy is reported. This accuracy is a measure of performance on '*unseen*' data.

4 GOTOV data

In this section, the data collection regarding the *Growing Old TOgether Validation* (GOTOV) study is presented, PART **A** (Fig. 1). The GOTOV study was designed to create a labelled dataset of physical activity, focused on older people. From this dataset, we can develop and compare multiple activity recognition models. To the extent of our knowledge, this GOTOV dataset is the first dataset with a focus on activity recognition of healthy older people using wearable sensors to be made publicly available.

4.1 Data collection protocol

The GOTOV collection took place at Leiden University Medical Center (LUMC), between February and May 2015. During this period, 35 (14 female, 21 male)
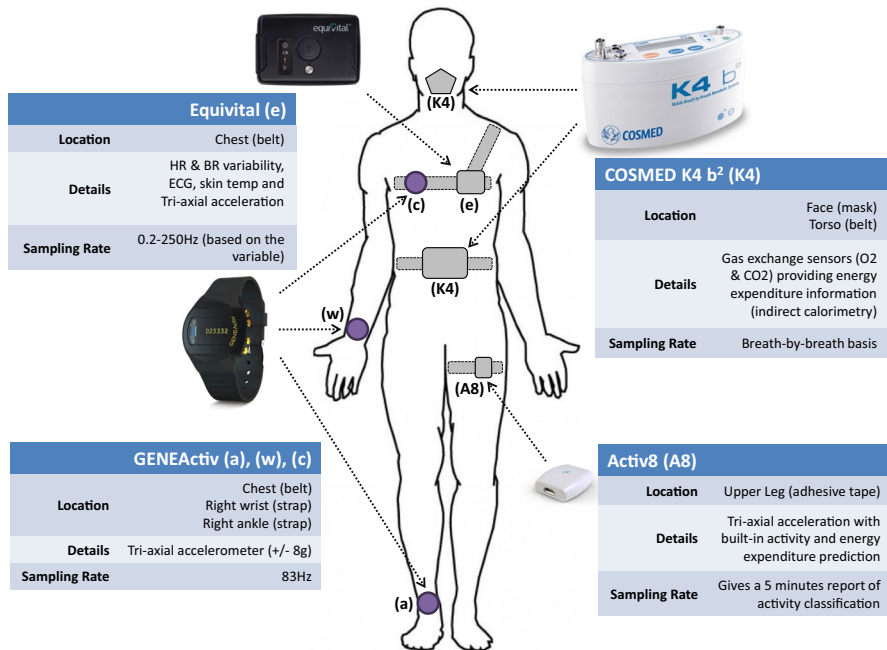
Fig. 3 GOTOV study devices and their body location

Table 1 The 16 activities protocol

Activity	Description	Duration in sec, expected (average)
Jumping	Lightly jump (synchronise sensors)	20 (20.1)
Standing	Get some rest between the two activities	60 (62)
Step	Stepping a step 20 times at participant's pace	60 (40.6)
Lying down left	Turn 90° to the left	180 (179.1)
Lying down right	Turn 90° to the right	180 (177.7)
Sitting sofa	Sit and watch TV (feet touch the ground)	180 (177)
Sitting couch	Sit and read a newspaper (feet on the Couch)	180 (180)
Sitting chair	Sit on an office chair word on a Computer	180 (173.1)
Walking stairs up	Ascend two flights of stairs	20 (17.3)
Washing dishes	Stand and wash dishes	180 (180)
Stacking shelves	Stack shelves with books	180 (180.6)
Vacuum cleaning	Perform some cleaning with a vacuum cleaner	180 (181.6)
Walking slow	Walk at a slow pace	300 (298.4)
Walking normal	Walk at a medium pace	300 (299)
Walking fast	Walk at a fast pace	300 (285.2)
Cycling	Cycle at a normal pace	900 (724.4)

individuals performed 16 everyday activities (see Fig. 2) while wearing 4 devices on 6 body locations (see Fig. 3). The set of 16 activities followed a specific, 90 min, protocol (see Table 1).

4.1.1 Participants selection

The individuals were selected via a paper advertisement and they had to meet the following criteria: (1) Be older than 60 years of age; (2) Have a BMI between 23 and 35 kg/m^2; (3) Not being restricted in their movements by health conditions; (4) Bring their own bicycle. The selected participants received €50 for their contribution to the GOTOV study and agreed to the use of recorded data for scientific purposes, in an anonymised manner.

4.1.2 Activities ontology

The set of activity classes to be performed were selected based on two main criteria. First, they had to represent the different aspects of everyday life, both in an indoor and outdoor environment. Second, and most importantly, they had to follow a hierarchical ontology with different levels of description. These levels allow us to generalise the set of activities depending on the activity detail needed in every AR model. We wanted to include in our protocol a variation of signals from similar activities that have:

- different patterns in sensor-readings because of body position/posture but are the same, like sitting on a sofa or chair is both sitting.
- different patterns in sensor-readings because they are performed at different intensities/paces per person, but they are the same activity, like walking.
- similar body position/posture and performed intensities and it is hard to distinguish them later, like different household activities.

In order to include all these different signals while training, we designed an activity ontology that hierarchically categorizes the activities. The design of this ontology tree, which can be seen in the two axes of our tree, see Fig. 1 of the paper, is based on two principals:

1. We should be able to classify activities at several intensities.
2. We should be able to classify activities at several levels of detail.

Based on these principals, we included 16 everyday activities, which were divided into 3 levels of description. In Fig. 2, we present the activities ontology and the level of description for each activity. The lower level of description (light blue) has been divided into low, mid and high-intensity. The second level from the top of Fig. 2 represents a set of 7 everyday activities (indoors and outdoors), the *General Activities* (see ontology Sect. 3). These activities were categorised in the 3 intensities according to the METs values per activity code from Ainsworth et al. (2011).

The set of indoor activities were mainly distributed under the low and mid-intensity with classes like lying down, sitting, standing and household chores, while the class of jumping belongs to high-intensity. On the other hand, the set of outdoor activities belong to the high-intensity group with classes like walking and cycling. Furthermore, we should point out that the outdoor activities were performed while the participants encountered different meteorological and traffic conditions such as crossroads or signs.

The third and highest level of activity description (dark blue) contains the more detailed activities, such as different sides of lying down (left and right), various types of sitting, multiple households and several paces of walking. In detail, the various sitting classes were divided into sitting on a chair (participants sit and work), sitting on a sofa (participants watched TV) and sitting on a couch (participants read a newspaper). The discrimination between sitting on Sofa and Couch is based on the fact that for Couch the participants placed their feet on it, while for the Sofa their feet were touching the ground. The multiple household chores were divided into washing dishes, stacking shelves and vacuum cleaning. Lastly, the different paces of walking were divided into walking with different paces (slow, normal, fast), walking stairs up and exercising with a step. We should point out that step and walking stairs up were performed indoors, while walking with different paces was performed outdoors.

4.1.3 The activities protocol

The 35 participants performed the set of 16 activities following a specific protocol. During that time, a researcher asked the participants to perform every activity during a time window. However, no other instructions or illustrations of the activities were given. After that, the researcher monitored the activities and recorded their start and end times. The order, duration (expected and average measured) and description of activities are presented in Table 1. A visual example of the procedure can be found in a recorded video.[1] Adding to that, every participant filled a physical activity questionnaire (IPAQ) (Craig et al. 2003), just before the monitoring protocol started.

Before every individual started the sequence of activities, there was a prior sensor calibration step of *COSMED* that took approximately 10–15 min. Following that, the individuals jumped lightly for 20 s, while waving their arms around. This procedure creates a recognisable pattern in the data, that is used to synchronise the sensors. Subsequently, the different activities were performed in the two environments, indoors and outdoors, with the outdoor activities performed in the immediate vicinity of the research building. Between every two activities (except lying down right and left), the participants were always standing for more than 1 min, providing this way with a clear demarcation to the signal data.

It can be observed that some activities were performed for a longer time compared to others (see Table 1). This was intended, in order to accommodate the complexity present in some of them (e.g. cycling). It should, also, be stated that due to

---

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

adverse weather conditions, some of the participants did not perform the outdoor activities (10 out of 35). Nevertheless, we decided to include these participants in our study in order to increase the number of instances for indoor activities. In the end, the data collection contained approximately 26 h of labelled activities.

4.1.4 Devices and body locations

The devices and their body locations were selected to serve Leiden biomedical studies mentioned before. The set of devices included both accelerometry sensors and sensors measuring participants' physical information, e.g. breathing rate (BR) and volume (VO_2, VCO_2) or heart rate (HR). The goal of using those sets of sensors is to observe diverse parameters for body motion or energy expenditure. In Fig. 3, the details of the devices and their respective body locations are presented. In more detail, we used the GENEActiv accelerometer on the ankle, wrist and chest, the Equivital measuring both accelerometry and other physical information on the chest, the *COSMED K4b*[2] measuring breath volume using a face mask and wearable unit on the torso, and *Activ8* which predicts energy expenditure from accelerometry data. If a device was severely limiting a participant's movement, it was removed (7 out of 35 participants did not use at least one device).

Since the focus of this study is on activity recognition models, the devices that would be used for our analysis are the combinations of GENEActiv accelerometers (*a,w,c* in Fig. 3) and Equivital (*e* in Fig. 3). On the other hand, data from *COSMED* (*K4* in Fig. 3) and *Activ8* (*A8* in Fig. 3) are intended for future research in order to develop energy expenditure models using as input also the activities predicted.

4.2 Publicly available dataset

One of the main contributions of this paper is to create a large publicly available dataset for physical activity recognition focused on older individuals, see (A) in Fig. 1. Motivated by that, the part of GOTOV study dataset, which is related to this activity recognition paper is made freely available in the *4TU.Center for research data* repository[2] (for DOI link see footnote 2). This AR dataset contains data from 2 out of 4 GOTOV study devices and includes 35 participants performing the 16 activities protocol. In detail, GENEActiv accelerometer on the ankle, wrist, chest (*a, w* and *c* in Fig. 3) and Equivital on the chest (*e* in Fig. 3) with both accelerometer and the accompanying physiological measurements data are publicly available. Having in mind the task of activity recognition, the resulting sensor set-ups and their abbreviations are presented in Table 2. Note that chest-Equivital (*ce*) sensor set-up combines e and Equivital, both placed on the chest (same body location).

The dataset made public is already a product of multiple pre-processing steps. Those steps assure that all the devices are synchronised and the activity labels are imputed according to protocol. Since every device uses different time formats, all

---

Table 2 Sensor set-ups abbreviation. (Color table online)

One-body location	Two-body locations	Three-body locations
a: ankle	*aw*: ankle-wrist	*awe*: ankle-wrist-Equivital
w: wrist	*ac*: ankle-chest	*awc*: ankle-wrist-chest
c: chest	*cw*: chest-wrist	*awce*: ankle-wrist-chest-Equivital
e: Equivital	*ae*: ankle-Equivital	
ce: chest-Equivital	*we*: wrist-Equivital	
	ace: ankle-chest-Equivital	
	wce: chest-wrist-Equivital	

The ankle, wrist and chest (*a,w,c* respectively) refer to GENEActiv sensor positioned at these body locations
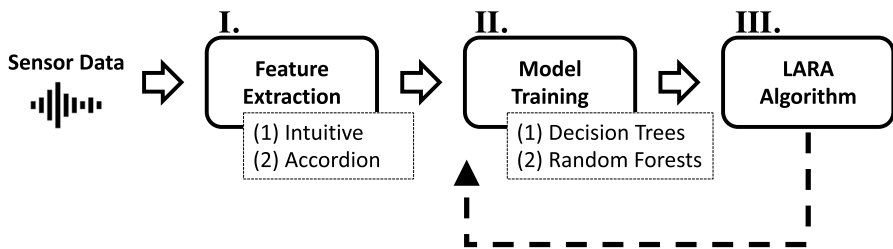

Fig. 4 Experiments workflow, where I refers to Sect. 5.1, II to Sect. 5.2, and III to Sects. 5.3 and 5.5

timestamps were converted to a UNIX format. UNIX timestamps, also known as POSIX time or epoch time, allow a synchronisation precision up to the millisecond. The resulting target (activities) and predictors (sensors) datasets are made publicly available for each participant.

5 Training activity recognition models

In this section, the training, evaluation and performance of the sensor set-ups are presented (B and C in Fig. 1). All the experiments were implemented in the R language, with external packages for machine learning and published algorithms for feature construction and selection. For all our experiments, we use the publicly available activity recognition dataset from the GOTOV study. For our study, we included the participants that had data for all three GENEActivs (ankle, chest, wrist) and Equivital device. For that reason, only 28 participants out of 35 were included in the current study, from which all the 28 performed the indoors activities and 18 the outdoors.

The structure of our experiments is presented in Fig. 4 and is as follows. First, we compare an intuitive feature construction (Baseline) with an automated one (Accordion), see I in Fig. 4. Second, the choice of classifier follows comparing Decision Trees (DT) and Random Forest (RF) (II in Fig. 4). Here, we analyse each activity precision and how it influences the total accuracy, computed by LOSO-CV, per sensor set-up. The Accordion, RF pipeline is extended with LARA to compare the classification power per sensor set-up, in terms of activity granularity (III in Fig. 4),

Table 3 Number of Accordion features per sensor variable and set-up. (Color table online)

	Set-up	One-body Location a	w	c	e	ce	Two-body Location aw	ac	ae	wc	we	ace	wce	Three-body Location awc	awe	awce
	Total N	140	224	164	186	138	148	119	64	163	163	96	143	118	85	222
	Avg Window	2.31	2.87	2.98	3.44	3.28	3.08	2.67	1.76	2.45	3.4	3.04	2.68	2.67	3.4	2.72
GENEActiv / ankle	X	58	–	–	–	–	34	27	12	–	–	16	–	25	14	49
	Y	51	–	–	–	–	35	28	9	–	–	19	–	24	17	39
	Z	31	–	–	–	–	20	20	8	–	–	8	–	21	3	21
GENEActiv / wrist	X	–	69	–	–	–	8	–	–	45	23	–	16	13	1	11
	Y	–	101	–	–	–	33	–	–	29	34	–	13	18	13	10
	Z	–	54	–	–	–	18	–	–	16	15	–	7	15	5	12
GENEActiv / chest	X	–	–	62	–	23	–	19	–	19	–	6	7	–	–	–
	Y	–	–	51	–	20	–	13	–	35	–	5	28	2	–	2
	Z	–	–	51	–	7	–	12	–	19	–	7	14	–	–	–
Equivital	Lat	–	–	–	29	14	–	–	6	–	17	9	6	–	3	17
	Ver	–	–	–	39	20	–	–	11	–	26	7	14	–	13	16
	Lon	–	–	–	40	25	–	–	9	–	17	10	19	–	7	15
	HR	–	–	–	23	3	–	–	1	–	5	–	1	–	1	2
	HRc	–	–	–	9	3	–	–	1	–	–	–	2	–	–	4
	BR	–	–	–	11	4	–	–	1	–	4	2	1	–	3	7
	BRc	–	–	–	4	3	–	–	1	–	3	1	–	–	–	3
	BW	–	–	–	2	1	–	–	–	–	3	1	3	–	1	2
	ECG	–	–	–	8	4	–	–	2	–	7	2	3	–	1	4
	IBi	–	–	–	7	4	–	–	–	–	1	–	1	–	–	2
	Temp	–	–	–	14	7	–	–	3	–	8	3	8	–	3	6

For set-up abbreviations, see Table 2. Equivital variables, *Lat, Ver, Lon* lateral, vertical, longitudinal acceleration, *HR, BR* heart and breath rate, *HRc, BRc* heart and breath rate confidence, *BW* breathing wave, *ECG* electrocardiogram, *IBi* inter-beat interval, *Temp* skin temperature

when accuracy of 90% is asked. The 90% threshold is chosen, having in mind future predictions in free-living data. This threshold is a parameter set in LARA in order to compare the performance of the different sensor set-ups. Nevertheless, it can be set differently depending on the confidence needed. Finally, we compare the performance of the developed activity recognition models per set-up.

5.1 Feature construction and selection

Here, we analyse the features constructed by the Baseline and Accordion. During the baseline experiment, we constructed 1 aggregated feature as a predictor, using the mean value of a fixed non-overlapping window of 1 s, for every sensor variable. Using these features, we trained a DT per sensor set-up and evaluated it with LOSO-CV. This experiment will allow us to compare between an intuitive and a data-driven feature construction strategy.

Our second step was to build and select features using the Accordion algorithm and train another DT. Accordion (Cachucho et al. 2014) is a machine learning wrapper that finds an optimised set of aggregate features, for each sensor set-up. This algorithm, given a maximum window size (5 s), constructs and selects aggregated features according to the *information gain* measure. The decision of limiting the window size to a maximum of 5 s is to enforce the trade-off between recognition speed and accuracy. In detail, if we use bigger windows in order to represent 1 s of activity it might be that there will be a loss of activity detail; since for example if we double the window size to 10 s, one is able in this time to alter between two different activities, e.g. sit-stand-sit and this in-between information might not be used correctly. Smaller windows will have the ability to react more quickly to changes in activity.

Table 3 presents the different number of features constructed by the Accordion algorithm per sensor (row) and sensor set-up (columns). In the table, the cells represent the number of features constructed and selected for each set-up. The highlighted in grey are the variables with the most features selected. In detail, Table 3 presents information about features selected (predictors) per data set-up, which is represented per column. The first two rows of the table show the total number of features selected per set-up (first row) and their average window size (2nd row). After these two rows, we present the number of features selected per devices' variable (predictor), and if a device was not included in this combination, we added a dash to its variables. In order to read this information correctly, we divided the table according to the two devices used, GENEActiv and Equivital. Then, since GENEActivs were used in 3 different body location we also divided this into three different groups, while for Equivital, placed only on the chest there is only one block of information. For example, if you want to see how many features were selected for ankle-wrist-equivital (*awe*) combination per variable, you should go to column *awe* and then one could observe that for GENEActiv on the ankle there were 14, 17 and 3 features selected per variable, with 17 being highlighted since it is the max of all the rows for column *awe*.

From Table 3, we can observe that whenever ankle is included in one data set-up, the most selected features are always from one of its variables (predictors). In other words, the ankle is the sensor with the most features selected whenever it is included in the sensors set-up. This indicates that an accelerometer on the lower limb is highly informative for ambulatory activities. Even when compared with physical measurements (breath rate, heart rate, etc.), an accelerometer is the more relevant for AR.

Every feature constructed and selected by Accordion is a representation of the variable used, followed by the aggregation function applied to it for the selected window. Some examples of selected features per sensor are *ankleY_SD_*39, *wristX_Mean_*74 and *SkinTemperature_SD_*6. In *ankleY_SD_*39 the *standard deviation* of ankleY (g force on the *y* axis) summarised for a window of 39 inputs is presented. Similarly, for the wrist, *wristX_Mean_*74 presents the *mean* of a window of 74 inputs on *X* axes acceleration variable. In the first two rows of the table, for each sensor set-up, we present the total number of selected features and their mean window size (in seconds).

In addition to Table 3, we plot the probability density function (PDF), using Kernel Density Estimation of feature windows that Accordion selected between 0 and 5 s per set-up, see Fig. 5. The 5 s here is the maximum window size given to Accordion. In the literature, the used windows sizes vary (Huynh and Schiele 2005; Banos et al. 2014, 2012). In Fig. 5, we observe two main modes of window sizes: 1–1.5 and 3.5–4 s. This is an important observation since the Accordion considers many candidate features and fixing the window size allows for a much quicker feature selection. Another interesting finding is that the graph does not peak around 5 s. This is important because if there would have been many features selected close to our specified maximum of 5 s, that would warrant further investigation of longer windows.
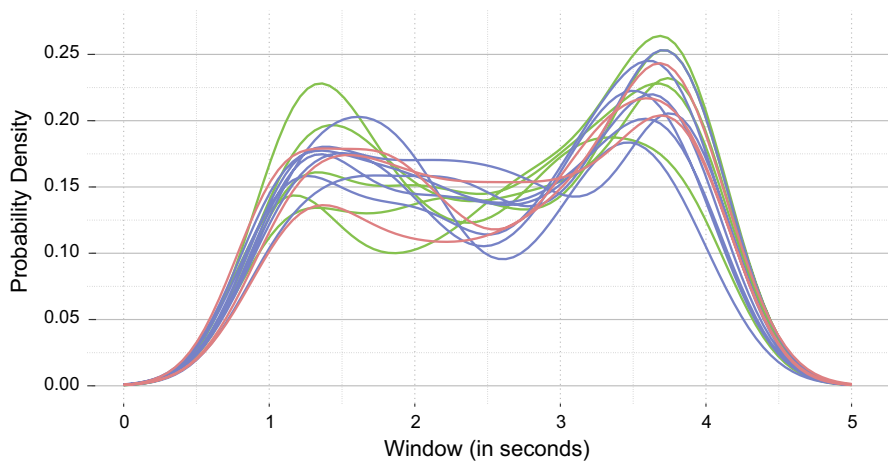
Fig. 5 The probability density function per window per data set-up. ■ one-body location, ■ : two-body locations, ■ : three-body locations. (Color figure online)

Table 4 Accuracies per experiment and sensor set-up (16-class classification). (Color table online)

Set-up		Baseline DT	Accordion DT	Accordion RF
Three	awce	49.4	73.5	82.6
Body	awe	56.9	78.7	86.8
Locations	awc	51	71.8	78.5
	wce	41.2	69.9	81.9
	ace	48.6	73.8	81.8
Two	we	39.6	73.8	81.9
Body	wc	42	74.3	80.6
Locations	ae	47.7	73.6	79.8
	ac	52.5	71.5	77.5
	aw	64.3	77.8	83
	ce	30.4	60.8	72.5
One	e	30.9	58.1	69.5
Body	c	31.4	57.4	62.6
Location	w	43.3	63.1	70.5
	a	50.3	66	71

The *a,w,c* refer to GENEActiv and *e* to Equivital, see Table 2

5.2 Evaluation of activity recognition models (C, Fig. 1)

To compare the two methods of data space representation, we trained a DT to classify each second of input to one of the 16 activities, using both Baseline and Accordion features for every sensor set-up. We evaluated the models using LOSO-CV and their performance is demonstrated in Table 4. For set-up abbreviations, see Table 2. As expected, Accordion outperforms the Baseline results. Compared to the Baseline,
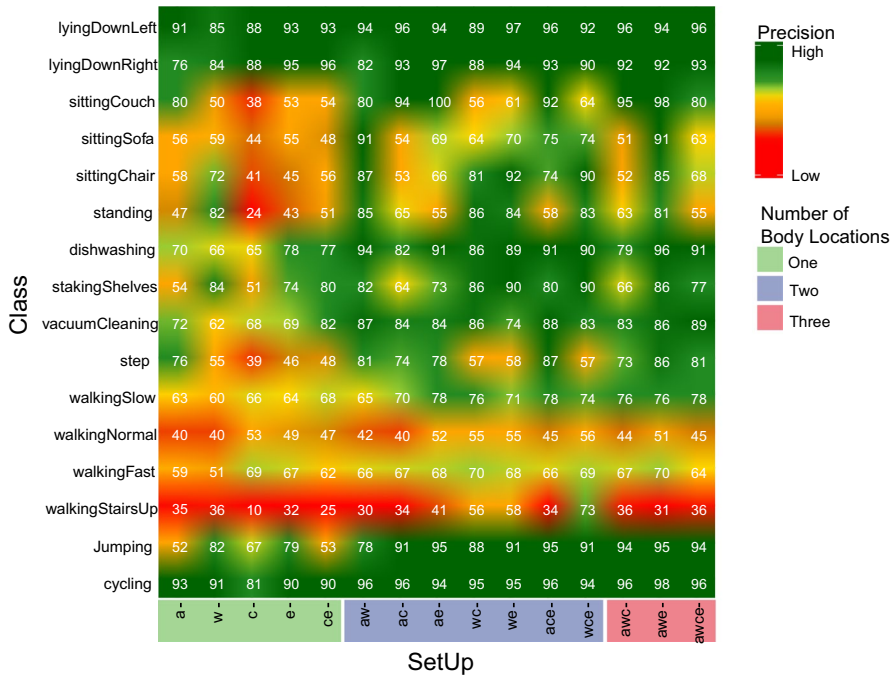
Fig. 6 Activities precision per sensor set-up. The *a,w,c* refer to GENEActiv and *e* to Equivital, see Table 2

Accordion takes into account the variations of the signal by using a set of aggregate functions and different window sizes. As a result, the models' accuracy increased by ≈ 20–30%.

From the machine learning perspective, there are numerous choices for classification algorithms. In our experiments, we compare DT (C4.5) with RF. Both DT and RF have been used extensively for activity recognition problems (Bao and Intille 2004; Reiss and Stricker 2012; Patel et al. 2009; Parkka et al. 2006; Gao et al. 2014; Sasaki et al. 2016). When learning RF using the Accordion features, we improved the activity recognition models for every sensor set-up. Compared with DT using the same features, the accuracy increased ≈ 6–12% (Table 4). Please note that using RF, all sensor set-ups that combine multiple body locations achieve an accuracy higher than 77%, where *aw* and *awe* have the higher accuracies of 83% and 86.8% respectively (16-class classification).

In Fig. 6, we display a heatmap of classification precision (%) per set-up. In other words, the percentage of an activity to be True Positive over all the true inputs of this activity per different sensor combination (*y*-axes). From this figure, it is clear that *cycling*, *lying down* and *jumping* are the activities with the highest precision. On the other hand, the poor performing activities are *standing* and the several types of *sitting* and *walking*. For these activities, we can make two main observations. First, we notice that their poor performance is only observed for one body location set-ups. This is represented by the orange-red area in the upper left corner of the

precision heatmap. This pattern indicates that, by measuring multiple body locations, one can improve the precision of the low and mid-intensity activities.

As for our second observation, we see that the various paces of *walking* are underperforming regardless of the sensor set-up. This is the horizontal orange-red area on the bottom of the precision heatmap. To know more about these two under-performance patterns, we have to examine the confusion between activity prediction. In order to do that, we examine the confusion matrices per set-up. In detail, we converted the confusion matrices counts into precision and error percentages per activity by normalising their results. These percentual results account for activity classes' imbalance. For practical reasons, in Fig. 7, we present two of the confusion matrices, one of the single (above) and one of the multiple body locations (below). The displayed confusion matrices highlight the percentage of confusion (light green to red) and precision (green to dark green) per class.

In most cases, adding an extra body location creates activity recognition models with higher class precision. In fact, when comparing single and multiple body locations, the sitting and household classes are distinguished more clearly. On the other hand, the various walking classes remain highly confused. In Fig. 7, one can see how confusion is distributed among the walking types. In fact, *walkingNormal* is confused both with *walkingSlow* and *walkingFast*. This is likely to be due to the fact that every individual has a different walking pace.

5.3 LARA implementation

To understand why our models do not reach the given LOSO-CV threshold accuracy of 90% we implemented the LARA algorithm as presented in Sect. 3.2.2. Here, we present an example of LARA's first iteration procedures and rationales for the algorithmic choices we did. The choices are related to the machine learning algorithm (RF) chosen for classification, the bottom-up merging procedure for classes of the same supra-class (parents) and the criteria to stop learning (90% threshold). Following Algorithm 1, we analyse the model's performance by the different activities.

As a result, when applying LARA, merging the walking classes to their parent class would be the first step in order for our models to achieve the given accuracy threshold. The improvement, in terms of accuracy, after LARA's first iteration with one walking class, is displayed in Table 5 (12-class classification). For set-up abbreviations, see Table 2. As expected, merging the walking classes resulted in a substantial increase in accuracy for all the models ($\approx 10\%$). The sensor set-ups that overcame the 90% threshold accuracy with the first merging, as seen in Table 5, are *aw* (93.9%), *ace* (90.2%), *wce* (90%) and *awe* (94.4%). Analysing the accuracies of different sensor set-ups (12-class classification), it seems that there are two groups of results: one around 80% and another around 90%. The first group is the one with sensors located only at a single body location, while the second is a combination of multiple. As for the top performers, the two models with the highest accuracy combine sensors on the ankle and wrist, *aw* with 93.9% accuracy or on the ankle, wrist
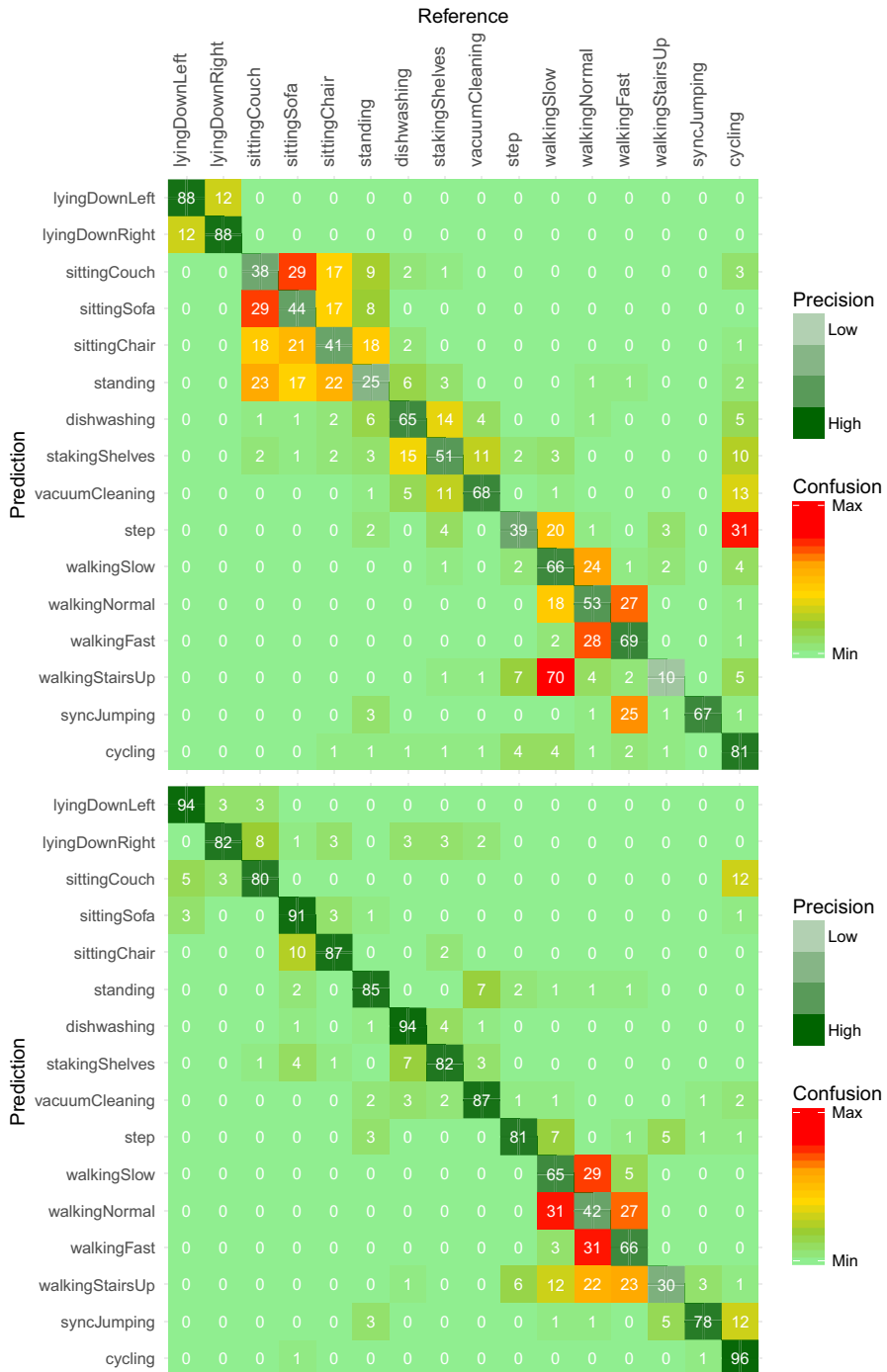
Fig. 7 Confusion matrices of chest (*c*, top) and ankle & wrist (*aw*, bottom) sensor set-ups. (Color figure online)

Table 5 Accuracies per merging using LARA algorithm. (Color table online)

	Set-up	16-class	12-class	10-class	8-class	7-class
Three	awce	82.6	89.1	91	-	-
Body	awe	86.8	94.4	-	-	-
Locations	awc	78.5	85.6	90	-	-
	wce	81.9	90	-	-	-
	ace	81.8	90.2	-	-	-
Two	we	81.9	88.6	92.8	-	-
Body	wc	80.6	86.9	92.4	-	-
Locations	ae	79.8	89.4	91.3	-	-
	ac	77.5	85.7	88.4	91.3	-
	aw	83	93.9	-	-	-
	ce	72.5	79.5	87.6	90.3	-
One	e	69.5	79	86.7	90	-
Body	c	62.6	69.8	78.7	85.4	85.7
Location	w	70.5	80.7	85.3	90.2	-
	a	71	81.8	84	90	-

Here *a,w,c* refer to GENEActiv and *e* to Equivital, see Table 2
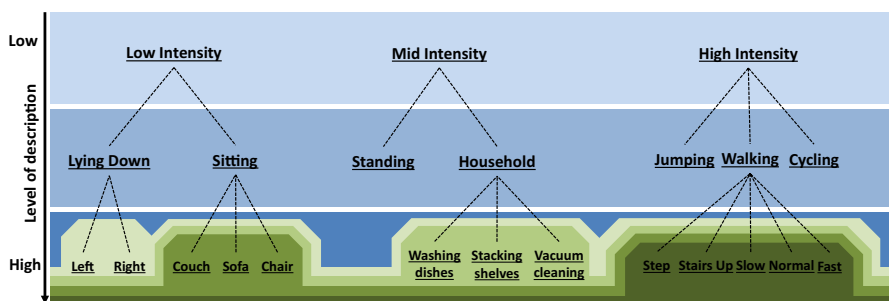
Fig. 8 Activity ontology with the steps of LARA merging. ■: 1st Iteration, merging walking paces for {*aw*, *ace*, *wce*, *awe*}. ■: 2nd Iteration, merging different sitting for {*ae*, *wc*, *we*, *awc*, *awce*}. ■: 3rd Iteration, merging household chores for {*a*, *w*, *e*, *ce*, *ac*}. ■: 4th Iteration, merging lyingDownLeft and lyingDownRight for {*c*}. For set-up abbreviations, see Table 2. (Color figure online)

and Equivital on the chest,*awe* with 94.4% accuracy. The difference in performance between these set-ups due to K–S test[3] is not significant.

Following the LARA procedure, in Fig. 8, the different merges per iteration and the sensor set-ups reaching the accuracy threshold are displayed. During the second iteration, the various sitting classes are merged into one class. Here, it can be observed that with a 10-class classification, one class for sitting and another one for walking, all the sensor set-ups using more than one body location reach an accuracy of 90%, except for the *ac*, Table 5. Note that, if a set-up reached the given threshold

---

[3] Kolmogorov–Smirnov test.

it was not trained in the next iteration. Further on, by the third iteration, the household chores collapse to one class (8-class classification) and all set-ups except the chest have an accuracy of more than 90%. Finally, it is clear that a chest accelerometer does not have the desired classification performance, since it does not reach the threshold value even when lying down classes are merged. Having in mind that the mid-level our P_{min} stopping criterion, the LARA algorithm stops and returns that a chest accelerometer is not enough to reach the desired classification accuracy (90%) for the minimum level of description.

5.4 Sensor set-ups

Examining the merging results of LARA algorithm in Fig. 8, we can decide which sensor location could be the most suitable for similar studies. As examples, we consider two different scenarios: the minimal sensor set-up (best body location) and the best performing set-up in terms of accuracy. While proposing different set-ups, we consider multiple perspectives such as the classification model performance and the comfort of the devices.

Starting with the minimal sensor location, when devices are placed only on one body location, ankle, wrist, Equivital (sensor strap placed on chest) and the combination of chest accelerometer and Equivital, the performance of these set-ups is similar, in reaching the 90% threshold accuracy only after LARA's third iteration (8-class classification). At this iteration, LARA suggests to merge the different kinds of walking, sitting and household activities. For most application needs, these levels of description are still quite acceptable when considering only one sensor location.

When more than one body location can be selected, adding to the ankle either wrist or chest (Equivital) produces very good solutions. This was already expected from the feature analysis done in Sect. 5.1. Considering also LARA's first iteration, combining ankle and wrist creates highly accurate and robust models, with high levels of description (12-class classification). In detail, the AR models combining ankle and wrist achieve 93.9% accuracy for *aw* and 94.4% for *awe*. However, please note that this difference is not significant.

On the other hand, it seems that more data from more body locations does not always result in better models. This can be observed in some of the AR models produced by LARA, such as set-ups with ankle, wrist and chest (*awc*) or even more clearly, ankle, wrist, chest and Equivital (*awce*). The model accuracies of *awc* and *awce* reach the 90% threshold only after LARA's third iteration (10-class classification).

5.5 Demonstration

Since the AR models intended to be used in studies with unlabeled accelerometer data, in this section (part **D**, Fig. 1), we demonstrate on how our 12-class-*aw* model performs on free-living activity data. In order to do that, we will first evaluate our model on a free-living dataset with labelled activities (**D-1**, Fig. 1). Then, we will
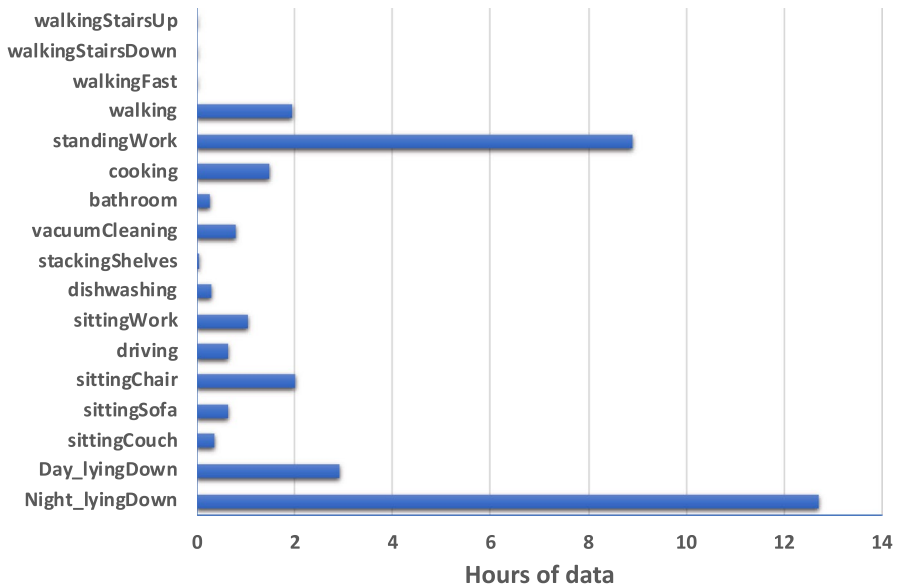
Fig. 9 Hours of activities reported during the free-living data collection

discuss how the LARA algorithm could be used in order to see which level of activity granularity to use when we try to predict the free-living activities. Finally, we will apply our model to one of our existing studies, the GOTO intervention study (van de Rest et al. 2016) (**D-2**, Fig. 1).

5.5.1 Free-living labelled data (part D, Fig. 1)

In order to validate our models in free-living data collections, there was a need for a free-living labelled data collection focused on older participants. Since there was no such collection freely available, we initiated our own pilot study. There are three main ways to perform such a data collection:

- Participants log their activities by recording times and descriptions.
- One or multiple researchers observe and label the activities performed.
- Stationary or body cameras record everyday life of participants, and later, one or multiple researchers label the activities with the use of the video.

For this study, we decided to perform a data collection using a combination of the above. We recruited two participants, one male, one female, that had demographics similar to the GOTO study in terms of BMI, age and health status. The participants had a researcher follow them during the day for at least 4 h on 2 different days, with one session taking place in the morning and one in the afternoon. The participants reported themselves their sleeping times and if they woke up during the night. During the day sessions, the researcher reported the activities performed using a mobile
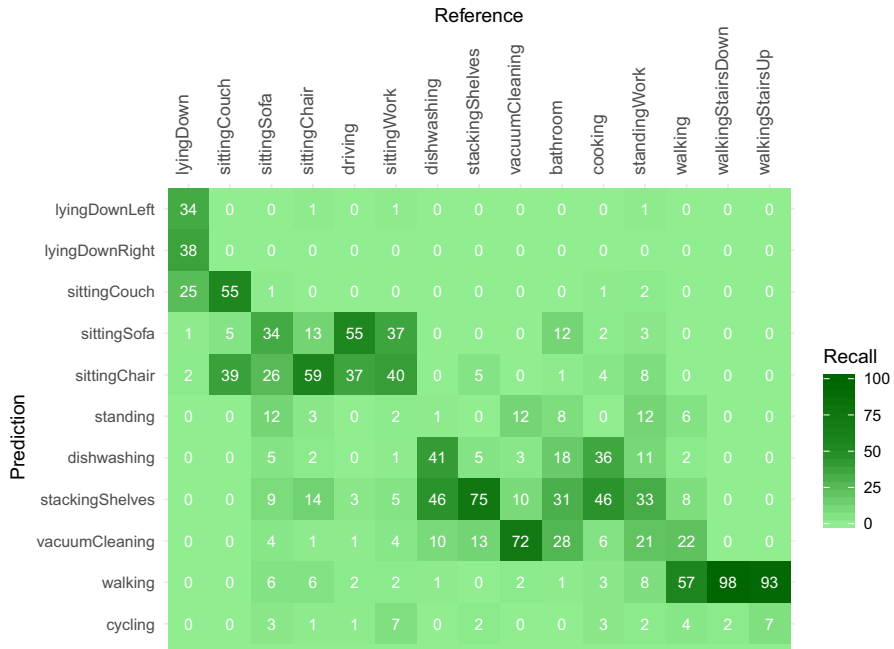
Fig. 10 12-class model results when plotted against the classes collected in free-living

app, *Atracker* (ATracker)[4] and was allowed to take pictures or videos of the activities performed if necessary. During the data collection, only two accelerometers were used, one on the wrist and one on the ankle, like in the GOTO study set-up (van de Rest et al. 2016). Both participants signed an informed consent stating that their activity data could be shared with our research team to validate the developed AR models.

In total, 32 h of free-living data were collected, 16 h of 20 daytime activities and approx. 16 h of sleeping. In Fig. 9, the total hours of 17 activities is presented since two general activity classes were included in our set, *sittingWork* and *standingWork*. These classes represent multiple too short or too general activities while sitting or standing (for example, gardening, watering plants, etc.). Note that the existing activities from GOTOV were employed, but also several new classes were introduced to fit activities not included in the validation study (e.g. driving).

Having the labels and the accelerometer data, we are able to apply and evaluate our GOTOV-developed AR models. Taking into account the fact that there are several new labels in the free-living activities that did not exist in the activity ontology, it is not possible to measure our model's performance with the use of accuracy

---

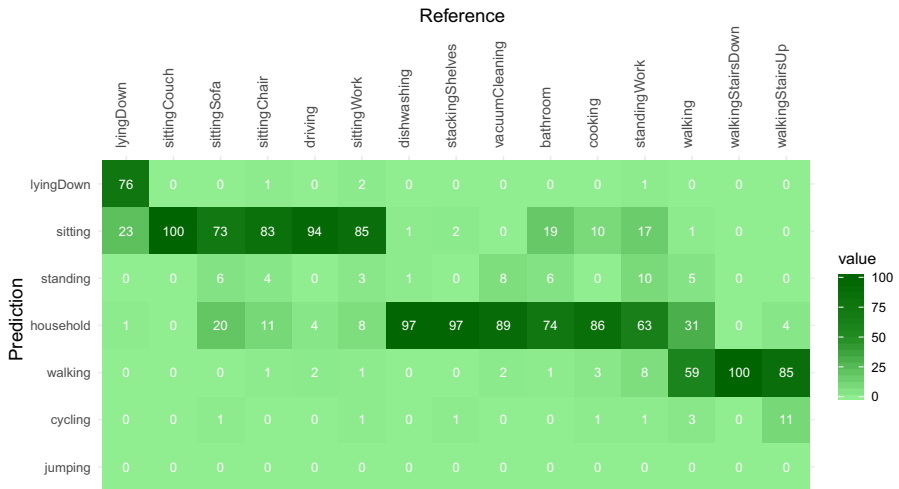[4] ATracker: Daily Task and Time Tracking, Time management app on iOS and Android, https://apps.apple.com/us/app/atracker-time-tracker/id522008611 and https://play.google.com/store/apps/details?id=com.wonderapps.ATracker&hl=en. Accessed date: 11 Sept 2019.

Prediction \ Reference	lyingDown	sittingCouch	sittingSofa	sittingChair	driving	sittingWork	dishwashing	stackingShelves	vacuumCleaning	bathroom	cooking	standingWork	walking	walkingStairsDown	walkingStairsUp
lyingDown	76	0	0	1	0	2	0	0	0	0	0	1	0	0	0
sitting	23	100	73	83	94	85	1	2	0	19	10	17	1	0	0
standing	0	0	6	4	0	3	1	0	8	6	0	10	5	0	0
household	1	0	20	11	4	8	97	97	89	74	86	63	31	0	4
walking	0	0	0	1	2	1	0	0	2	1	3	8	59	100	85
cycling	0	0	1	0	0	1	0	1	0	0	1	1	3	0	11
jumping	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

value: 100, 75, 50, 25, 0

Fig. 11 7-class model results when plotted against the classes collected in free-living

metrics. This is because generally, there is no one-to-one activity comparison between the activities that the models were trained and the free-living activities. Consider Fig. 10, in order to evaluate whether the predicted activities are similar to the real free-living ones.

In Fig. 10, we present how the free-living activities are predicted using the 12-class *aw* model presented earlier. For example, we observe that 34% of the *lyingDown* instances are predicted as *lyingDownLeft*, 38% as *lyingDownRight* and 25% as *sittingCouch*. These values could be thought of as a Recall since we compare the number of predicted true positives over the sum of predicted true positives and true negatives. This can be seen with classes that exist in our protocol. For example, *sittingCouch*, where 55% of its instances are predicted correctly while 39% are predicted as *sittingChair*.

Keeping that in mind, we can identify four blocks of confusion based on these values: (a) the lying down, (b) the sitting activities, (c) the activities that include standing and main use of hand and (d) the walkings. From these blocks, we can observe that activities that were not included in our hierarchy are predicted as activities that make use of the same body parts. For instance, *driving* is predicted mainly as *sittingSofa* (55%) or *sittingChair* (37%), while *cooking* 36% of the time is predicted as *dishwashing* and 46% as *stackingShelves*. Both of these make sense, since driving is a sedentary activity while cooking is a household activity where hands are used while standing.

In conclusion, we understand from this analysis that predicting free-living activities with laboratory-trained models is a challenging task. Having this in mind, we decided to try reducing the activity detail level by using the LARA algorithm.

	Precision	Recall	F1	Balanced accuracy
lyingDown	0.99	0.76	0.86	0.88
sitting	0.43	0.85	0.57	0.83
standing	0	–	–	–
household	0.87	0.69	0.77	0.82
walking	0.58	0.60	0.59	0.79
jumping	0	–	–	–
cycling	0	–	–	–

Table 6 Activities in both training and free-living labelled data with their performance metrics

5.5.2 LARA implementation on free-living labelled dataset

Applying the LARA algorithm to our 12-class free-living predictions requires a small modification. This is because in the laboratory environment, we measured confusion as the *false discovery rate*, since we could compare class on a one-to-one basis using *precision*, while now we will need to compute confusion as the *false-negative rate* based on *recall*.

The LARA algorithm followed three iterations, one per parent activity of the mid-level activity description (see Fig. 2) except walking, which was already merged before. During the three iterations, the first activities merged were the household ones, while the sittings followed. In the last iteration, *lyingDownLeft* and *lying-DownRight* were merged to *lyingDown*. The result of the produced 7-class model is demonstrated in Fig. 11.

Similar to Fig. 10, Fig. 11 represents how every true class is predicted. Here, we can see clearer the four groups of activities. The first is the *lyingDown* activities. The second is all the *sitting* activities, the third all the activities performed while standing and the last is the different types *walking*. The two classes the higher confusion rate (with several other classes) are *standingWork* and *walking*. In more detail, *standingWork* is confused with multiple classes since this activity includes multiple smaller ones like gardening, plant watering, desk cleaning, house maintenance, etc. On the other hand, *walking* is mainly confused with household activities. This confusion mostly results from *vacuumCleaning* (see Fig. 10), since this activity also includes steps and not only standing and using the hands like dishwashing. Another interesting point is that there is a confusion between *sitting* and *lyingDown*, which is mainly because of *sittingCouch* activity (see Fig. 10). This sitting is different from the other ones since the feet of the subjects do not touch the ground but they are positioned on the couch or the coffee table. This results in a similar body position with lying down from the view of ankle accelerometer signal.

Taking into account these results, we are confident that when predicting in free-living conditions with our 7-class model, we can identify and classify most of the activity patterns included in these 7 groups. As a result, we can conclude to an updated activity ontology that includes activities that were not part of the GOTOV protocol, based on their main confusion with one of our 7-classes. Thus, the *driving*
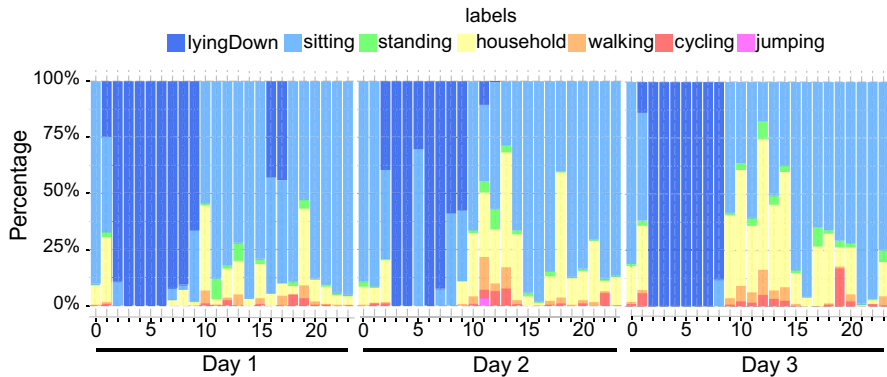
Fig. 12 Activity prediction per day of a GOTO participant

and *sittingWork* can be added as children of *sitting*, the *bathroom*, *cooking* and *standingWork* as children of *household* and the *walkingStairsDown* under *walking*.

With such an ontology, we can have one-to-one comparison of the classes and present the main metrics, see Table 6. Here, we see that *lyingDown* is predicted with 99% precision, *household* with 87%, while *sitting* and *walking* have lower performances with 43% and 58% accordingly. In detail, as discussed, the main confusion of *sitting* is with *lyingDown* due to the *sittingCouch* class, while for *walking* the main confusion is with *household*. This is expected since the activities represented as *household* are consisted by standing, main move of upper limp but also short steps, e.g. *vacuumCleaning* and *cooking*. In Table 6, we present the 7 activities with their performance metrics. Have in mind, that there was no *standing* still, *cycling* and *jumping* performed during our free-living data collection.

Concluding, if we want to examine how our model performs in terms of accuracy, we will need to use a metric that cope with an imbalanced classification problem (see Fig. 9), such a metric is the *Balanced Accuracy* (Velez et al. 2008).

$$Balanced\ Accuracy = (Sensitivity + Specificity)/2.$$

The reported average balanced accuracy of our model is 74%. From the analysis above, it is clear that the confusion mostly stems from activities that make use of the same body parts. So, the value of 74% accuracy is not that low under this consideration. Adding to that, when compared with previous work of *Sasaki* (Sasaki et al. 2016), where they also developed RF models trained in laboratory condition and tested on free-living data, they report an accuracy of 49–55% for a 5-class model, while our approach results in higher accuracy (74%) and resolution (7-classes). These improved results are likely the result of our models being based on two sensor instead of one for Sasaki et al.

5.5.3 GOTO application (part D-2, Fig. 1)

Since the AR models intended to be used in studies with unlabelled accelerometer data, we perform a short demonstration of the 7-class *aw* model in the unlabelled

free-living data of the GOTO study (part **D-2**, Fig. 1). The dataset used is part of the GOTO intervention study (van de Rest et al. 2016) where two accelerometers were used to monitor 160 participants. As part of our demonstration, we selected one participant who used both ankle and wrist GENEActiv. In Fig. 12, we present the activity predictions of *aw*-model built on the 7-class scenario. This demonstration shows how our model, in the future, can be used to develop and aggregate participants activity profiles.

From Fig. 12 we can see a clear pattern of active and non-active behaviour during day and night, respectively. In detail, during the day, the two dominant activity classes are sitting and household activities (doing something while sitting or standing), as expected. On the other hand, there is a high percentage of sitting during the night, which as discussed a bit earlier could be lying down with a body position similar to sittingCouch. Furthermore, we can see an increase of activity levels over the three days, with day 1 having lower activity levels than 2 and day 3 more than 2. Nevertheless, this variation of activity patterns should also be examined in more detail, something that is beyond the scope of this paper.

6 Conclusions and future work

In this paper, we present the *Growing Old TOgether Validation* (GOTOV) study for monitoring physical activity, in order to serve multiple mobility and healthy ageing studies in older adults. Throughout the paper, our main goals involve four main topics. First, we generated a new activity recognition dataset, using wearable sensors in a population of individuals over 60 years old. This dataset boasts a hierarchy of activity classes, contains both indoor and outdoor activities and is now made publicly available. Second, we developed LARA, a method to learn robust and accurate AR models. Third, we delivered a sensor set-up analysis focusing on which body locations or combinations are the most efficient to monitor and predict physical activity. Fourth, we provide an AR model that can be used in free-living data in order to recognise general physical activity patterns that can be associated with physiological health parameters. Working towards these goals, our conclusions are presented in the following paragraphs.

The GOTOV dataset is special for a number of reasons. To begin with, it is the first publicly available dataset that focuses on older individuals (aged over 60). The focus until now has been mostly on a younger population (mid-twenties), which is known to faster adopt wearable technology. Additionally, it is a reasonably large data collection, with 35 participants performing 16 everyday-life activities while wearing multiple devices on three different body locations, both accelerometers and physiological sensors.

Furthermore, the data collection protocol has been designed and executed to mimic free-living challenges, being collected both indoors and outdoors. The different activity classes allow several levels of activity detail, which follow a specific activity ontology. Hence, the dataset allows creating different AR models, depending on the desired level of activity detail and the available sensor set-up on focused.

Here, we would like also to discuss some limitations of our data collection and as a result our modelling. First, we should note that the market of wearables has developed impressively in the last years. This implies that there could be smarter systems to measure physical activities when compared to our sensors. Second, during the data collection some activity classes were performed for longer and some other for shorter periods of time (see Table 1 for details) which resulted in a somewhat unbalanced dataset. Adding to that, it would be better in the future to perform the same activities more than once per participants and possibly without following a specific order, in order to capture more signal variation from every activity. Third, after our demonstration, we realise that from one side it is not feasible to include all of the free-living activities in a training set, but on the other side having an ontology where the description can be adjusted could be really useful in order to capture the more general activities, as presented with the 7-class model analysis. However, it is reasonable that the use of free-living data for training would be ideal, but such data are really challenging to collect.

Nevertheless, our data collection includes a wide range of activities with different granularity and of sufficient size. As a result, the GOTOV dataset as well as the developed AR models can be related to sensor data from more than 500 participants ($\approx$ 3 Tb), which can be combined with high-value health data from the Leiden Biomedical Ageing Studies (see "Appendix 2").

In detail, about the AR models, our main goal was to be robust with high accuracy. Using the LOSO-CV evaluation technique, we designed a wrapper algorithm that reduces the number of classes belonging to the same supra-class, until an accuracy threshold is achieved. We combined this method with automated feature construction and selection and a well-known learning algorithm of ensemble models. As a result, for each subset of sensors (set-up), we output a robust Random Forest model with different capacity of activity recognition. Here, we should point out that an algorithm like LARA, that merges classes to improve accuracy levels, even though we use it for such a purpose, reduces the classification power. On the other hand, it is a useful tool for comparing the activity granularity a given set-up requires to achieve a specific accuracy. The choice for Random Forests was made in order to be able to investigate the contribution of body location and different devices for AR. Still, we would like to stress that our pipeline is independent of the classification algorithm used. Concluding, the combination of Accordion, RF and LARA using an activity ontology is a valuable pipeline for activity recognition research.

Having multiple sensor locations, we were able to indicate which is the optimal sensor set-up for activity recognition. From all the models built, our best AR models are a combination of merging the three paces of walking (12-class classification) and the use of ankle, wrist accelerometers and Equivital. This model achieves an accuracy of 94.4%, while it contains a considerably high level of activity detail. Nevertheless, leaving out the Equivital, the combination of ankle wrist does not have a much lower accuracy (93.9%).

Comparing the minimal set-ups (so one device only), for an 8-class classification problem, we see that ankle, wrist and Equivital (accelerometer and physiological measurements) have similar accuracies. Nevertheless, considering the market availability and comfort, it the wrist accelerometer (see Table 5, 8-class, with 90.2%

accuracy) seems to be the most recommendable minimal set-up. The wrist is a practical solution in terms of body location, with multiple market solutions (smart watches) being available.

6.1 Future work

Now that we have shown that the *aw* AR model is able to quite accurately recognise activities in 7 classes in free-living data, we will proceed applying this model to Leiden biomedical datasets (see "Appendix 1"). These datasets have been collected to study the association of frequency, duration and intensity of activities, with health parameters in older people. We aim to develop general physical activity patterns and associate them with physiological parameters of health like blood pressure, lipid and glucose levels. Furthermore, since it is clear that training on free-living data could improve the accuracy, this motivates us to consider a full-scale labelled free-living data collection.

Except for the activity recognition part, we would like to observe the activity patterns per person and predict the energy expenditure of these. To further investigate that and to what extent the activity intensity levels are metabolically effective, in the GOTOV data collection, already the energy expenditure (using the COSMED device) has been included. In this manner, we will attempt to estimate the intensities of activities performed by each person in our Leiden biomedical studies, predict the energy expenditure and correlate them with the collected health parameters.

In the long run, in the Leiden biomedical studies, our goal would be to observe among older adults which range and intensity of physical activities improve parameters of metabolic health, general health and well-being. These relations, then, can 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.

Appendix 1: Table of abbreviations across the paper

See Table 7.

Table 7 Abbreviations across the paper in alphabetical order

Abbreviations	
a	GENEActiv ankle
A8	Activ8
AR	Activity recognition
BMI	Body Mass Index
BR	Breath rate
BRc	Breath rate confidence
BW	Breath wave
c	GENEActiv chest
DT	Decision tree
e	Equivital
ECG	Electrocardiogram
GOTO	Growing old together study
GOTOV	Growing old together validation study
HR	Heart rate
HRc	Heart rate confidence
IBI	Inter beat interval
ICG	Intra-group confusion
K-S	Kolmogorov–Smirnov test
K4	COSMED K4b^2
LARA	Learning activity recognition accurately
Lat	Equivital lateral acceleration
Lon	Equivital longitudinal acceleration
LOSO-CV	Leave one subject out cross-validation
MET	Metabolic equivalent of task
PDF	Probability density function
RF	Random Forest
Temp	Skin temperature
Ver	Equivital vertical acceleration
w	GENEActiv wrist

Appendix 2: Further data: Leiden biomedical ageing studies

All Leiden Biomedical Ageing Studies were approved by the Medical Ethical Committee of the Leiden University Medical Centre and was performed according to the Helsinki declaration. All participants gave written informed consent for participation.

Leiden Longevity Study

The Leiden Longevity Study (LLS) (Westendorp et al. 2009) was initiated to identify the genetic and environmental determinants of human longevity and healthy ageing. The LLS participants have been recruited between 2003 and 2006 and the complete study consists of 421 long-lived families including 944 siblings of 90 years and older, their middle-aged offspring ($N = 1674$, mean age $= 60$ years) as a group of predisposed to become long-lived and age healthily, and spouses/partners ($N = 744$, mean age 60 years) of their offspring as representatives of the general population. We found that middle-aged members of long-lived families have a lower prevalence of hypertension, cardiovascular disease and type 2 diabetes as compared to the controls. In addition, we found that members of long-lived families are less likely to be a carrier of the genetic APOEε4 allele and more likely to be a carrier of the genetic APOEε2 allele (Deelen 2014).

All participants of the LLS have been followed for their vital status, and in 2017 we obtained data on present and developed morbidities of the offspring and control generation, as well as their medication use. In these members of long-lived families and the controls, we have collected 1 week of free-living wearable data (GENEActiv), which we would like to annotate the frequency and intensity of different activities and relate that to health parameters such as clinical blood values, anthropometrics, blood pressure, blood metabolite and gene expression levels, and brain MRI images that in the past have been collected in the same participants.

Growing Old TOgether study

The Growing Old TOgether (GOTO) study (van de Rest et al. 2016) is a cohort of LLS where we applied in older people a 13-weeks lifestyle program, with a target of 12.5% caloric restriction and 12.5% increase in energy expenditure through an increase in physical activity. For people in their 40s and 50s, lifestyle programs have been shown to improve metabolic health. For older adults, however, it is not clear whether these programs are equally healthy. For the GOTO study, we recruited, in 2012, participants from the LLS, i.e. couples of members of long-lived families and their current partners (controls). In total, 164 older adults (mean age $= 63.2$ years; BMI$= 23–35$ kg/m^2) participated. The mean weight loss of baseline weight was 4.2% (SE $= 2.8$%), while many other parameters of metabolic health, such as blood pressure, and thyroid, glucose and lipid metabolism improved significantly. Many metabolites levels changed in a direction previously associated with a low risk of

type 2 diabetes and cardiovascular disease and partially independently of weight loss. In the GOTO study, we collected wearable sensor data (GENEActiv) 1 week before the start of the intervention and the 13th week of the intervention. We would like to annotate the available wearable sensor data to frequency and intensity of different activities and relate that to (the change in) health parameters, for example clinical blood values, anthropometrics, blood pressure, blood metabolite and gene expression levels, and brain MRI images that have been collected in the same participants.

Active and Healthy Ageing study

In 2011, in the Active and Healthy Ageing (AGO) study (Wijsman et al. 2013), we applied in older people a 12-weeks web-based physical activity program (DirectLife, Philips, Consumer Lifestyle, Amsterdam) directed at increasing daily physical activity. The DirectLife program consists of three elements: (1) an accelerometer-based activity monitor, (2) a personal website and (3) a personal e-coach, who provides regular updates of the individual's physical activity status by email and gives advice to increase physical activities. In total, 235 older adults (mean age 64.8 years) participated in the AGO study. Body composition (mean weight loss was 1.5 kg) and measures of glucose metabolism improved. In the AGO study, we collected wearable sensor data (GENEActiv) 1 week before the start of the intervention and the 12th week of the intervention. We would like to annotate the available wearable sensor data to frequency and intensity of different activities and relate that to (the change in) health parameters, for example clinical blood values, anthropometrics, blood pressure and blood metabolite that have been collected in the same participants.

Switchbox study

The Switchbox study (Jansen 2015) has been initiated to study endocrine and metabolic regulation in relation to health in old age and participants were recruited from the LLS. Between January 2012 and April 2013, 135 members of long-lived families and controls from the LLS were recruited and selected based on previous obtained information and were middle-aged (55–77 year) with a body mass index (BMI) between 19 and 33 kg/m^2 to participate.

In the Switchbox study, we collected wearable sensor data (GENEActiv and Equivital belt) during 1 week. We would like to annotate the available wearable sensor data to frequency and intensity of different activities and relate that to health parameters, for example clinical blood values, anthropometrics, continuous physiological measurements, blood pressure and blood metabolite that have been collected in the same participants.

References

Ainsworth, B.E., Haskell, W.L., Herrmann, S.D., Meckes Jr., N., Bassett, D.R., Tudor-Lock, C., Greer, J.L., Vezina, J., Whitt-Glove, M.C., Leon, A.S.: 2011 Compendium of physical activities: a second update of codes and MET values. Aging Clin. Exo Res. **43**, 1575–1581 (2011)

Altun, K., Barshan, B.: Human activity recognition using inertial/magnetic sensor units. In: Proceedings of the International Workshop on Human Behavior Understanding, pp. 38–51 (2010)

Attal, F., Ferhat, S., Dedabrishvili, M., Chamroukhi, F., Oukhellou, L., Amirat, Y.: Physical human activity recognition using wearable sensors. Sensors **15**, 31314–31338 (2015)

Banos, O., Galvez, J.M., Damas, M., Pomares, H., Rojas, I.: Centinela: a human activity recognition system based on acceleration and vital sign data. Pervasive Mobile Comput. **8**, 717–729 (2012)

Banos, O., Galvez, J.M., Damas, M., Pomares, H., Rojas, I.: Window size impact in human activity recognition. Sensors **14**, 6474–6499 (2014)

Bao, L., Intille, S.: Activity recognition from user-annotated acceleration data. In: Pervasive Computing: Second International Conference (PERVASIVE), pp. 1–17 (2004)

Bifet, A.: Adaptive Learning and Mining for Data Streams and frequent Patterns, Ph.D. thesis. Universitat Politècnica de Catalunya (2009)

Breiman, L.: Random forests. Mach. Learn. **4**, 5–32 (2001)

Brush, A., Krumm, J., Scott, J., Saponas, S.: Recognizing activities from mobile sensor data: challenges and opportunities. In: Proceedings of Ubicomp (2011)

Cachucho, R., Meeng, M., Vespier, U., Nijssen, S., Knobbe, A.: Mining multivariate time series with mixed sampling rates. In: Proceedings of Ubicomp, pp. 413–423 (2014)

Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S.T., Tröster, G., Millán, J., Roggen, D.: The Opportunity challenge: a benchmark database for on-body sensor-based activity recognition. Pattern Recogn. Lett. **34**, 2033–2042 (2013)

Chen, L., Nugent, C.: Ontology-based activity recognition in intelligent pervasive environments. Med. Biol. Eng. Comput. **5**, 410–430 (2009)

Craig, C.L., Marshall, A.L., Sjöström, M., Bauman, A.E., Booth, M.L., Ainsworth, B.E., et al.: International physical activity questionnaire: 12-country reliability and validity. Med. Sci. Sports Exerc. **35**(8), 1381–95 (2003)

Deelen, J.: Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. Hum. Mol. Genet. **23**, 4420–4432 (2014)

Del Rosario, M.B., Wang, K., Wang, J., Liu, Y., Brodie, M., Delbaere, K., Lovell, N.H., Lord, S.R., Redmond, S.J.: A comparison of activity classification in younger and older cohorts using a smartphone. Physiol Meas. **35**, 2269–286 (2014)

Duong, T.V., Bui, H.H., Phung, D.Q., Venkatesh, S.: Activity recognition and abnormality detection with the switching hidden semi-Markov model. In: Procedings 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 838–845 (2005)

Dutta, A., Ma, O., Buman, M.P., Bliss, D.W.: Learning approach for classification of GENEActiv accelerometer data for unique activity identification. In: 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), pp. 359–364 (2016)

Gao, L., Bourke, A., Nelson, J.: Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems. Med. Eng. Phys. **36**, 779–785 (2014)

Huynh, T., Schiele, B.: Analyzing features for activity recognition. In: Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-Aware Services: Usages and Technologies (sOc-EUSAI '05), pp. 159–163 (2005)

Jansen, S.W.: Human longevity is characterised by high thyroid stimulating hormone secretion without altered energy metabolism. J. Sci. Rep. **5**, 11525 (2015)

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

Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutor. **15**, 1192–1209 (2013)

Martin, K.R., Koster, A., et al.: Changes in daily activity patterns with age in US men and women: National Health and Nutrition Examination Survey 2003–04 and 2005–06. J. Am. Geriatr. Soc. **62**, 1263–1271 (2014)

Martin, P.E., Rothstein, D.E., Larish, D.D.: Effects of age and physical activity status on the speed-aerobic demand relationship of walking. J. Appl. Physiol. **73**, 200–206 (1992)

Moshtaghi, M., Zukerman, I., Russell, R.A.: Statistical models for unobtrusively detecting abnormal periods of inactivity in older adults. User Model. User-Adapt. Interact. **25**, 231–265 (2015)

Murphy, S.L.: Review of physical activity measurement using accelerometers in older adults: considerations for research design and conduct. Prev. Med. **48**, 108–114 (2009)

Parkka, J., Ermes, M., Korpipaa, P., Mantyjarvi, J., Peltola, J., Korhonen, I.: Activity classification using realistic data from wearable sensors. IEEE Trans. Inf. Technol. Biomed. **10**, 119–128 (2006)

Patel, S., Mancinelli, C., Healey, J., Moy, M., Bonato, P.: Using wearable sensors to monitor physical activities of patients with COPD: a comparison of classifier performance. In: Proceedings Sixth International Workshop on Wearable and Implantable Body Sensor Networks, pp. 234–239 (2009)

Ranasinghe, S., Al Machot, F., Mayr, H.C.: A review on applications of activity recognition systems with regard to performance and evaluation. Int. J. Distrib. Sens. Netw. (2016). https://doi.org/10.1177/1550147716665520

Rashidi, P., Mihailidis, A.: A survey on ambient-assisted living tools for older adults. J. Biomed. Heath Inf. **17**, 579–590 (2013)

Reiss, A., Stricker, D.: Creating and benchmarking a new dataset for physical activity monitoring. In: Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments, pp. 40–48 (2012)

Rejeski, W.J., Marsh, A.P., Brubaker, P.H., Buman, M., Fielding, R.A., Hire, D., Manini, T., Rego, A., Miller, M.E.: Analysis and interpretation of accelerometry data in older adults: the LIFE study. J. Gerontol. Ser. A **71**, 521–528 (2016)

Riboni, D., Bettini, C.: OWL 2 modeling and reasoning with complex human activities. Pervasive Mobile Comput. **7**, 379–395 (2011)

Saez, Y., Baldominos, A., Isasi, P.: A comparison study of classifier algorithms for cross-person physical activity recognition. Sensors **17**, 66 (2016)

Sallis, J.F.: Age-related decline in physical activity: a synthesis of human and animal studies. Med. Sci. Sports Exerc. **32**, 1598–1600 (2000)

Sasaki, J.E., Hickey, A., Staudenmayer, J., John, D., Kent, J.A., Freedson, P.S.: Performance of activity classification algorithms in free-living older adults. Int. J. Web Inf. Syst. **48**, 941–950 (2016)

Schrack, J.A., Cooper, R., Koster, A., Shiroma, E.J., Murabito, J.M., Rejeski, W.J., Ferrucci, L., Harris, T.B.: Assessing daily physical activity in older adults: unraveling the complexity of monitors, measures, and methods. J. Gerontol. Ser. A **71**, 1039–1048 (2016)

Schrack, J.A., Leroux, A., Fleg, J.L., Zipunnikov, V., Simonsick, E.M., Studenski, S.A., Crainiceanu, C., Ferrucci, L.: Using heart rate and accelerometry to define quantity and intensity of physical activity in older adults. J. Gerontol. Ser. A **73**, 668–675 (2018)

Shiroma, E.J., Schrack, J.A., Harris, T.B.: Accelerating accelerometer research in aging. J. Gerontol. Ser. A **73**, 619–621 (2018)

Tedesco, S., Barton, J., O'Flynn, B.: A Review of activity trackers for senior citizens: research perspectives. Commercial Landscape and the Role of the Insurance Industry. J. Sens. **17**, 1277 (2017)

van de Rest, O., Schutte, B.A.M., Deelen, J., et al.: Metabolic effects of a 13-weeks lifestyle intervention in older adults: The Growing Old Together Study. Aging **8**, 111–124 (2016)

Velez, D.R., et al.: A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet. Epidemiol. **4**, 306 (2008)

Westendorp, R.G., et al.: Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: the Leiden Longevity Study. J. Am. Geriatr. Soc. **57**, 1634–1637 (2009)

Wijsman, C.A., et al.: Effects of a web-based intervention on physical activity and metabolism in older adults: randomized controlled trial. J. Med. Int. Res. **15**, e233 (2013)

Stylianos Paraschiakos is a data scientist/Ph.D. candidate in the Department of Biomedical Data Sciences, section Molecular Epidemiology of the Leiden University Medical Center in Leiden, the

Netherlands. He received his master in Computer Science and Advanced Data Analytics at the Leiden Institute of Advanced Computer Science. His work focuses on healthy ageing and how the physical activity of older people is linked to parameters of metabolic health and biological age.

Ricardo Cachucho Ph.D., is a data scientist and lead for Benelux and Nordics AI solutions at Pegasystems Inc. His research area is time series machine learning techniques with a focus on complex sensor networks. Cachucho's Ph.D. thesis is focused on the development of new machine learning algorithms that can capture classification patterns from multivariate sensor networks with mixed sampling rates.

Matthijs Moed is a senior advisor at SURF, the collaborative organisation for ICT in Dutch education and research. His work focuses primarily on building scalable data processing solutions with streaming data and big data technologies. Previously, he has worked as a bioinformatician in the field of molecular epidemiology for the investigation into human longevity and as a data scientist in the domain of medical image processing.

Diana van Heemst Ph.D., is an associate professor at in the Department of Internal Medicine, section Gerontology and Geriatrics of the Leiden University Medical Center (LUMC), the Netherlands. Her research area is the pathophysiology of human ageing. She integrates different approaches and methodologies to unravel the causal pathways that underlie the pathophysiology of human ageing in her research. These range from epidemiological data analyses in large cohorts and biobanks to hypothesis-driven in depth diagnostic studies, including randomised controlled trials. A central focus of her research is determining the role of endocrine and metabolic factors in the pathophysiology of human ageing and age-related diseases.

Simon Mooijaart MD, Ph.D., is an internist-geriatrician at the Leiden University Medical Center (LUMC), The Netherlands. His research area is Evidence-based medicine for older patients with the aim to improve the quality of healthcare for older patients, by generating, implementing and disseminating scientific evidence tailored for the older patient.

Eline P. Slagboom is head of the section of Molecular Epidemiology within the Department of Biomedical Data Sciences of the Leiden University Medical Center, the Netherlands. In addition, she is Fellow at the Max Planck Institute for Biology of Ageing in Cologne, Germany. She is chair of the Dutch Society for Research on Ageing (DuSRA) and board member of BBMRI-NL (Biobanking and Biomolecular Resources Research Infrastructure) steering national studies in metabolomics. The studies of her group focus on biomarkers and causal pathways for longevity, healthy ageing and metabolic health. Her research into stimulation of healthy ageing is focused on lifestyle intervention studies in humans.

Arno Knobbe is the head of the Data Mining & Sports group at Universiteit Leiden. His work is focused on the development of new data science techniques and the application of these on non-trivial, data-rich domains, for example in sports and health. His group has worked with a range of top sports teams, providing them with data-driven insights on training and tactics. Knobbe is also the director of the Sport Data Center, the consortium of Dutch universities that is creating the Sport Data Valley, which aims to become the national data facility for sports, both elite and amateur. The data science applications in health are focused on lifestyle monitoring and healthy ageing, by means of wearables.

Marian Beekman is a senior scientist in the Department of Biomedical Data Sciences, section Molecular Epidemiology of the Leiden University Medical Center, the Netherlands. She received her master in Biomedical Sciences at the Leiden University Medical Center for which she also received her Ph.D. in 2004. Since then she is investigating the determinants of human longevity and healthy ageing by studying long-lived families in transdisciplinary collaborations like Life Science, Epidemiology, Bioinformatics, Historical Databases, Gerontology and Health. Her ultimate aim is to prevent age-related disease, compress morbidity and extend human health span.

Affiliations

Stylianos Paraschiakos[1,2] (ORCID) **· Ricardo Cachucho**[1,2] **· Matthijs Moed**[1] **·
Diana van Heemst**[1] **· Simon Mooijaart**[1] **· Eline P. Slagboom**[1] **· Arno Knobbe**[2] **·
Marian Beekman**[1]

✉ Stylianos Paraschiakos
  s.paraschiakos@lumc.nl

[1] Leiden University Medical Center, Leiden, The Netherlands

[2] Leiden Institute of Advanced Computer Science, Leiden, The Netherlands