Discussion Paper

Comparing activity trackers to investigate physical activity: balancing quality, costs, and usability

Sigrid van Hoek[1]
Dick Windmeijer[1]
Annemieke Luiten[1]
John Bolte[2]
Barry Schouten[1]

[1]Statistics Netherlands [2]The Hague University of Applied Sciences

February 21, 2022

Table of Contents

Summary

This paper reports on the first stage of a research project[1] that aims to incorporate objective measures of physical activity into health and lifestyle surveys. Physical activity is typically measured with questionnaires that are known to have measurement issues, and specifically, overestimate the amount of physical activity of the population.

In a lab setting, 40 participants wore four different sensors on five different body parts, while performing various activities (sitting, standing, stepping with two intensities, bicycling with two intensities, walking stairs and jumping). During the first four activities, energy expenditure was measured by monitoring heart rate and the gas volume of in- and expired O_2 and CO_2. Participants subsequently wore two sensor systems (the ActivPAL on the thigh and the UKK on the waist) for a week. They also kept a diary keeping track of their physical activities, work and travel hours.

Machine learning algorithms were trained with different methods to determine which sensor and which method was best able to differentiate the various activities and the intensity with which they were performed. It was found that the ActivPAL had the highest overall accuracy, possibly because the data generated on the upper tigh seems to be best distinguishing between different types of activities and therefore led to the highest accuracy. Accuracy could be slightly increased by including measures of heartrate. For recognizing intensity, three different measures were compared: allocation of MET values to activities (used by ActivPAL), median absolute deviation, and heart rate. It turns out that each method has merits and disadvantages, but median absolute deviation seems to be the most promising metric. The search for the best method of gauging intensity is still ongoing.

Subsequently, the algorithms developed for the lab data were used to determine physical activity in the week people wore the devices during their everyday activities. It quickly turned out that the models are far from ready to be used on free living data. Two approaches are suggested to remedy this: additional research with meticulously labelled free living data, e.g., by combining a Time Use Survey with accelerometer measurements. The second is to focus on better determining intensity of movement, e.g., with the help of unsupervised pattern recognition techniques.

Accuracy was but one of the requirements for choosing a sensor system for subsequent research and ultimate implementation of sensor measurement in health surveys. Sensor position on the body, wearability, costs, usability, flexibility of analysis, response, and adherence to protocol equally determine the choice for a sensor. Also from these additional points of view, the activPAL is our sensor of choice.

[1] The authors would like to thank Reinier Bikker, Guido van den Heuvel, Jeffrey Hoogland and Bob Lodder for their effort to clean and standardize the data and perform explorative data analysis. We would also like to thank Mark Schauwen and the Educational Programme Motion Technology for facilitating the use of the laboratory and for the technical support. Thanks to THUAS students Adnan Akbas, Colin Werkhoven, Ali Safdari, Mark Boon and Matthew Turkenburg for their work on predicting activities in the free living week and to Utrecht University student Jasper Vogelzang for his work on Hidden Markov models to predict transitions between activities.

1 Introduction

Every year adults are asked to fill a short questionnaire to assess their physical activity and sedentary behaviour. Traditionally, Statistics Netherlands and the National Institute for Public Health and the Environment of the Netherlands (RIVM) measure the health of the Dutch population via a Short QUestionnaire to ASsess Health-enhancing physical activity (SQUASH), (Wendel-Vos et al., 2003). Respondents fill in the frequency and duration of several activities like walking, cycling and sports. Statistics Netherlands publishes each year about the population's adherence to Physical Activity Guidelines of the National Health Council. Physical activity is defined as any bodily movement produced by the skeletal muscle that results in energy expenditure (Prince et al., 2008). However, it is well-known that self-reported survey data are prone to measurement errors and lack of representativeness (Ferrari et al., 2007; Fruin and Rankin, 2004; Helmerhorst et al., 2012; Sallis and Saelens, 2000; Shephard, 2003; Welk et al., 2014; Wijndaele et al., 2015). Self report also suffers from reporting bias, e.g. due to social desirability or inaccurate memory (Helmerhorst et al., 2012; Prince et al., 2008; Sallis and Saelens, 2000; Welk et al., 2014). Respondents often underestimate the time spent in light intensity activities and overestimate high intensity activities (Nicolaou et al., 2016).

Activity monitor devices allow to objectively measure intensity, duration and patterns of activity (Troiano and Freedson, 2010). Tri-axial accelerometers can provide information about physically active and inactive periods (Ward et al., 2005). Accelerometers measure movements accurately and are widely available at a low cost (Esliger and Tremblay, 2007). Next to acceleration, activity monitor devices may measure for example heart rate, oxygen intake and counting steps.

However, research indicates that people are substantially more willing to fill in a questionnaire than to wear an accelerometer (e.g., Kraakman, 2021, Toepoel et al., 2021). It is unclear to what extent respondents are willing to wear activity trackers for research purposes, or to what extent activity trackers can replace surveys from a data quality perspective. A research project was started to address these questions (Schouten and Voermans, 2019). Specifically, we study whether it is possible to develop a sensor system and algorithm that are able to replace the SQUASH questionnaire.

This paper describes the first step in this research project: the search for a sensor system that is accurately and validly capable to recognise physical activity and its intensity. To determine accuracy and validity, a ground truth is necessary. The intensity of activities can be bench-marked with respiratory gas exhalation and heart rate. For recognition of physical activity we used an observation method, where participants performed a number of prescribed activities (cycling, walking, running, jumping, standing, climbing stairs, sitting) with five motion sensors at different positions of the body; and a heart rate sensor and VOX sensor attached to their body.

A small scale observation study (n=40) was performed in the motion technology laboratory of the Hague University of Applied Sciences (THUAS). In addition, participants wore two sensors at home for a week, while keeping a diary of their physical activities during that week. We will refer to the data that is generated during and outside the laboratory as respectively labdata and weekdata.

In this paper we describe the estimation of intensity and activity recognition. The report, particularly the activity classification in Section 4, builds on findings from a previous sub-project concerning the preliminary investigation of the data (Bikker et al., 2020). The result of this preliminary analysis was a model that classified the activities fairly well, with a few exceptions. Based on these analyses, it was determined that the activPAL in combination with a heart rate measurement classified the activities most accurately. In this report additional analyses with alternative machine learning models are used to study if the preliminary results are still upheld.

1.1 Research questions

CBS wants to determine whether sensor-based measurements can replace or supplement the SQUASH survey. The research consists of two aspects: classification of physical activity and estimating intensity. There is no true value for intensity during the free living period, therefore, we will only apply a classification algorithm based on the weekdata. The following research questions will be assessed in this report:

RQ1. How can we estimate intensity based on the lab- and weekdata?
RQ2. How well can machine learning models classify activities based on labdata?
RQ3. How well can machine learning models classify activities based on weekdata?

The feasibility of using accelerometers in large scale surveillance depends on a number of aspects. The three research questions above are concerned with the quality of measurement and the possibility to gauge the required variables from measures and algorithms. For the present pilot we chose to provide all participants with the same high quality research grade measurement devices, to be worn in the same position and on the same body parts. Other options may be feasible too, for example using relatively cheap commercial devices (like fitbits) that may already be in the participants' possession or are provided by the research institute. A pilot into the feasibility of using participants' own accelerometers is described in Kraakman (2021). Decisions on the type of measurement device to use, how long to measure, where to measure, all impact the balance between quality, costs and participant burden. Sample persons have to be willing and able to wear the meters, and the lowest possible respondent burden should be striven for. Respondent burden was not the main research topic for the pilot described here, nevertheless, some information is available that is addressed in the fourth research question:

RQ4. How do the sensors used in the free living period compare in terms of user friendliness?

The international physical activity norm states that an adult needs at least 150 minutes of minimally moderate physical activity per week and needs to perform muscle and bone-strengthening activities at least twice a week. The intensity of physical activity is expressed as Metabolic Equivalent of Task (MET). Moderate physical activity is defined as having a MET value of at least 3. Although MET value is calculated by dividing the oxygen uptake by the mass of a person in kilogram times 3.5 (Mortazavi et al., 2013), in practice the calculation of the physical activity norm is based on assigning a fixed MET value per type of activity. We compare the (pre-)calculated MET value with oxygen intake and other measures of intensity like the Mean Amplitude Deviation (MAD) that describes the typical distance of data points to their mean to answer RQ1.

In the SQUASH questionnaire, respondents are asked to estimate the number of minutes generally spent per week on a number of activities that typically take up most of people's time: working, travelling, biking for work and leisure, walking for work and for leisure, doing sports, gardening, doing household chores. These activities are assigned a MET value to be able to gauge the intensity of movement, for the classification in light, moderate and intensive physical activity. When using an accelerometer, only the direction and amplitude of the movement is measured.

Whether this movement is ’walking’ of ’bicycling’ needs to be recognised from the pattern of movement. Some accelerometers can for example use information on the inclination of the device to distinguish biking from walking, for example because the have a gyroscope (like the x-IMU’s) or by calculating inclination from the angles between the x-, y-, and z- vectors (like the activPAL software does). The step frequency distinguishes walking from running. Algorithms then translate these patterns into activity and intensity classification. An important consideration for the present research project is to what extent we will be able to develop these algorithms for intensity and activity recognition, and to what extent respondents will still have to provide this context. A subsequent question is which kind of activities need to be recognized in order to be able to provide the statistics needed.

Because the physical activity norm states the required number of active minutes per week, a week is also the measurement unit of time chosen in this study, as well as in most other studies with accelerometers, although shorter periods of four days may also render valid estimates of physical activity (Edwardson et al., 2017). Edwardson et al. (2017) gives an overview of studies using the activPAL. In 71% of studies, the activPAL was worn for 7 days. Ideally, the number of days of measurement are determined by the required reliability, which may be different for different types of behaviour (sitting, standing, walking, etc.) and different populations. Most likely, there will be variation over days of the week, over weeks and over seasons in the amount of physical activity. Questionnaires tackle this variability by asking for a subjective estimation of the mean number of minutes over longer periods of time. Longer measurement periods with accelerometers could theoretically inform on within respondent variation and could potentially better inform machine learning models. However, better measurement needs to be balanced with higher respondent burden and diminished adherence. Skin irritation may be the result when wearing the activPAL for longer periods, for example. There is also the practical consideration that there is a finite amount of recording capacity on the device. Edwardson et al. (2017) conclude that, pending better recommendations, the measurement period should be at least 7 days but ideally up to the 14 day limitation of the activPAL.

The second and third research questions will be answered by training supervised machine learning algorithms and evaluating their performance. The machine learning algorithms classify the sensor measurements into periods of activities like running, standing, and cycling. The performance of the models is evaluated via the logbook that is kept during the laboratory session. We compare the performance of the sensors that provided raw accelerometer data to answer RQ2. The logbook is not available for the weekdata. Therefore, we compare the classification of the weekdata of RQ3 with the SQUASH diaries and the predictions of the software of the ActivPAL sensor.

The structure of this report is the following. In Section 2 we will describe the sensors, lab sessions and diary in more detail. This section also contains information about data cleaning procedures. In Section 3 we evaluate three methods to estimate the intensity of physical activities. Section 4 describes the data analysis leading to the choice for one of the sensors (the activPAL) for subsequent studies in the research program. Neural networks are trained to classify activities by accelerometer data from the laboratory sessions. We also investigate how the models that were developed for recognising laboratory activities extrapolate to the free living period. Section 5 describes the lab and free living sensors from the viewpoint of user friendliness and other practical considerations. We finish the report with conclusions and recommendations for future work in Section 6.

2 Methods

Participants came to the THUAS motion technology laboratory, where they first signed a consent form and completed a questionnaire about their exercise behaviour (the SQUASH questionnaire). Two lab assistants weighed and measured the participants and attached five different sensors to various part of the body. Subsequently, participants performed a series of prescribed activities. During a number of activities, a breath gas analysis was done to measure energy expenditure. A detailed description of tasks is available in Luiten and Voermans (2019a) and Luiten and Voermans (2019b). Following the lab session, participants wore two sensors, the activPAL on the left upper thigh and the UKK at the back attached on a belt around the waist, at home for a week. During that week they filled in a diary in which they noted the times for getting up, working, travelling (plus mode), exercising, taking off the sensor(s), and going to bed. Some of the participants had the free living week prior to the lab session. They received the ActivPAL and UKK through the mail. A short manual assisted in mounting the sensors properly. Participants received €50 for their assistance.

2.1 Sample, sensors and activities

A convenience sample of 40 participants took part in the research in the laboratory. These were recruited among CBS employees via an e-mail, students of THUAS via posters at the THUAS, and participants of the CBS Health Survey who had indicated that they would be available for subsequent research again. An attempt was made to compose four homogeneous groups for this training set: half men, half women and half physically active, half inactive. The 'active' were not supposed to be a top athlete, and the 'inactive' were not supposed to be completely inert. The level of fitness, together with other variables such as age, correlate for example to the maximal exercise heart rate (Londeree and Moeschberger, 1982). Whether people were active or inactive was determined on the basis of three questions that participants had to answer: the number of minutes per week walking, cycling and exercising. The derivation of 'active' on the basis of these questions was mainly done on the basis of the number of minutes of exercise. The non-active participants walked an average of 63 minutes per day, cycled 29 minutes and played sports for 12 minutes, according to their answers to these questions. The active participants walked 66 minutes, cycled 32 minutes and exercised 45 minutes. Whether a person was classified as active or inactive influenced the setting of the activities that participants did in the laboratory session.

We were able to analyse the labdata of ActivPAL for 36 participants [2]. Table 2.1 shows the prior distribution of gender and activity of the 36 analysed participants. 55% of them were active, but being active was not equally distributed across the sexes: the male respondents were more active than female respondents. 9 respondents are members of a sports association and 13 respondents did fitness at home or at the gym. The weight of respondents ranges from 50.4 to 100.3 with a mean of 74.6 kg. The length of respondents ranges from 157.0 to 194.5 with a mean of 179.0 centimetre. The age of respondents ranges from 18 to 71 with an average of 35 years.

The following sensors were attached to participants during the laboratory sessions:

- The Vyntus CPX system. Vyntus allows to determine a subjects' metabolic response while

[2] The data quality of one respondent was problematic, the data of two respondents was incomplete, one respondent did not perform jumping in the laboratory.

Table 2.1 Gauged activity level and gender of participants

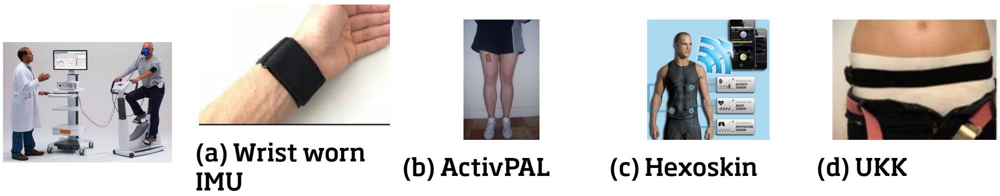
General activity level	Sex	Number of respondents
Active	Men	15
	Women	5
Inactive	Men	5
	Women	11

exercising. Through a face mask or mouth-piece the gas volume of in- and expired O_2 and CO_2 are collected. From the breathing volume and the differences between inspiration and expiration O_2 and CO_2 concentration the oxygen uptake and the CO_2 production are calculated by the software. The Vyntus was attached to a ergo-meter and treadmill for cycling and walking activities (see below for a description of the activities performed).

- The activPAL3 ™(PAL Technologies Ltd., Glasgow, UK) is a small and slim device that directly measures the postural aspect of sedentary and active behaviour. It is mounted on the upper thigh with a medicinal patch. Via proprietary algorithms information about thigh position and acceleration are used to determine body posture (i.e., sitting/lying and upright), stepping, and stepping speed (cadence), from which energy expenditure is inferred indirectly. The sampling frequency for acceleration data was 20 Hz. ActivPAL3 has 8bit output, so the measured accelerometer data is within the range $[0, 255]$.
- Two IMUs (Inertial Measurement Units), one mounted on the right wrist, one on the right tibia. The version used was the X-IMU (x-io Technologies Limited, Bristol, UK). The IMU contains a magnetometer, accelerometer, barometer and gyroscope. The sample frequency for acceleration data was 512 Hz. The X-IMU outputs 16bit acceleration data with a range of 8g.
- The Hexoskin smart shirt that measures heart rate, respiratory rate, minute ventilation, step count, and energy expenditure. The sampling frequency for acceleration data was 64 Hz.
- The UKK RM42 (UKK, Tampere, Finland) accelerometer. Raw acceleration data were collected at a 100Hz sampling rate with 13-bit A/D conversion of the ±16g range and analysed with a custom-written MATLAB script for mean amplitude deviation (MAD) in non-overlapping 5s epochs (Vähä-Ypyä, Vasankari, Husu, Suni, et al., 2015). Twelve consecutive 5s epochs were averaged to produce minute-by-minute MADs. Previously defined and validated cut-points of 16.7mg (mili-acceleration caused by gravity), 91mg and 414mg were used to categorise the minutes into separate sedentary time and light, moderate and vigorous intensity physical activity (Vähä-Ypyä, Vasankari, Husu, Mänttäri, et al., 2015; Vähä-Ypyä, Vasankari, Husu, Suni, et al., 2015). The UKK was worn in an elastic band on the hip during daytime and transferred to a wrist worn elastic band during sleeping hours.

The sensors and position on the body are shown in Figure 2.1. The same IMU sensor is worn on the ankle and wrist. ActivPAL, Hexoskin, two IMUs and UKK contain accelerometers. We will refer to the unprocessed tri-axial accelerometer data as raw data. ActivPAL provides data in raw and pre-classified (i.e., sitting/lying, standing, stepping) form. A summary file provides for example the number of steps, the proportion of time spend lying and the estimated MET. The pre-classification algorithms are not public, and the taxonomy of activities differs from our seven activities. The data of the UKK sensor are already processed into e.g., average X-, Y-, and Z acceleration at a frequency of approximately 1 Hz and includes the variable MAD (Mean Amplitude Deviation).

Measurements started with a synchronisation jump by the participant. This made it easier to



(a) Wrist worn IMU

(b) ActivPAL

(c) Hexoskin

(d) UKK

Figure 2.1 Sensors

superimpose the start of the measurements for all the sensors. The activity classes in which gas volume was also measured were:

- cycling on an ergo-meter, moderately intensive at 70 watts for untrained and 100 watts for trained participants. Revolution was kept at 60 per minute. We will refer to this activity as cycling light.
- cycling, intensive, at 80 and 120 watts for untrained and trained participants respectively, both also at 60 revolutions per minute. We will refer to this activity as cycling heavy.
- walking on the treadmill (4 km/h for untrained and 5 km/h for trained participants),
- running on the treadmill (8 km/h for untrained and 10 km/h for trained participants). Participants who could not run for five minutes could stop earlier or could run at a slower speed; it appeared that even for trained women 10 km/h was quite intensive.
- standing,
- sitting.

The activities were determined as representative activities that people perform daily. The intensity of walking and cycling for trained and untrained participants was based on literature and the experiences of the HUAS. The order of the activities was alternated among participants. However, running always followed walking, and cycling intensive always followed cycling moderate. Each activity lasted 5 minutes. Between walking / running and between cycling there was a short interval where the Vyntus needed to be hooked to the other device. Otherwise, the interval between activities was generally shorter than a minute, but could vary between respondents as a result of various circumstances (e.g., adjustment of sensors that came loose, blue tooth connections that were disrupted, etc.).

The activity classes in which no gas volume was measured were:

- Jumping: 5 jumps with 10 second intervals.
- Climbing stairs (up and down twice on the stairs between two levels of the laboratory).

Figure 2.2 shows the time series of accelerometer data of several sensors during the laboratory session for one of the respondents. UKK data of the activity sitting are missing for 10 respondents, and standing for 1 respondent. Although time shifts are already applied to the data, the start- and end-time of the logbook is sometimes still a bit off. For example, it is remarkable that VO$_2$ of running does not immediately increase.

The participants wore two sensors, ActivPAL and UKK, during a 7 day period while they continued with their normal life. During this week the participants kept an activity diary in which they indicated when they worked, when they cycled, when they travelled, when they sported, which sport they performed and when they slept. The diary also contains the start and end times of non-wear periods. Note that the activities noted in the diary were partly more extensive than those measured in the lab (e.g., travelling), and partly less extensive (e.g., we did not ask participants to note when they were sitting or standing). The diaries were mostly filled in at the

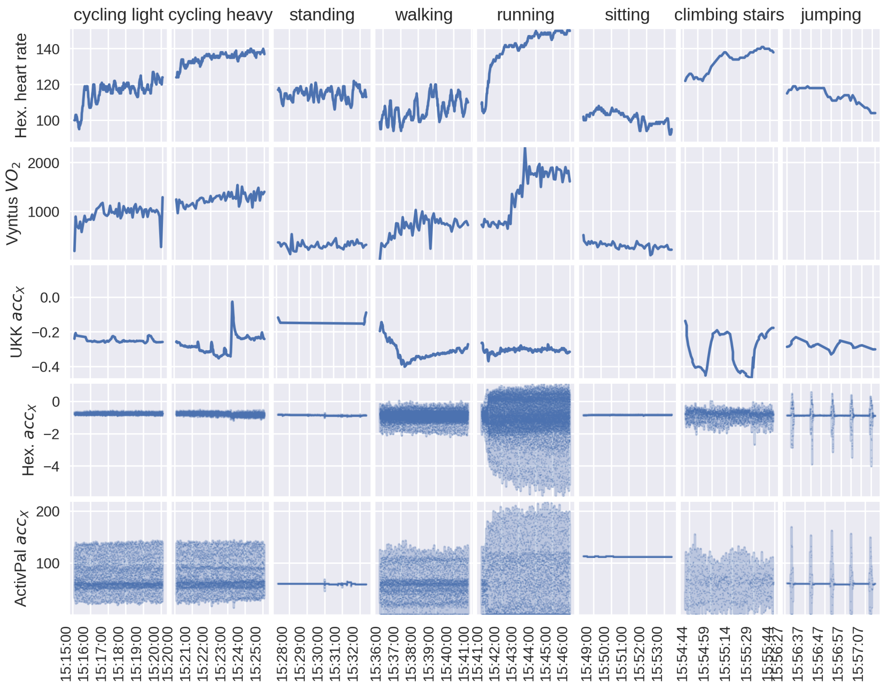


Figure 2.2 Measurements of different sensors during the lab sessions of one respondent after applying time corrections.

conclusion of a day, or even the following day, leading to inaccuracies. Some obvious mistakes like wrong dates and wrong hours were removed. For example, the date was be outside the measurement period or the hours indicated that the activity longer than reasonable. We also encountered some unlikely, but possible activities such as high intensity activities that could not be distinguished by accelerometer data and activPAL's MET values. The dairies were cleaned to a certain degree with the help of the activPAL data where the start and finish of certain activities could easily be ascertained. The diaries could not be cleaned thoroughly however.

The diary contains the variables 'wake up' and 'sleeping' to indicate sleeping. Sleeping is clearly visible in the accelerometer data of Figure 2.3. Other diary activities such as working and travelling are harder to see in the data. The working hours, 06:30 until 19:30, overlap with travelling, indicating that the diaries are not as accurate as the laboratory activities.

Moreover, respondents also completed the SQUASH questionnaire during the week. The activities in the completed questionnaire are mapped to intensities called Metabolic Equivalent of Task (MET) by assigning a fixed MET value per activity. The MET values are taken from Ainsworth's compendium of physical activity Ainsworth et al., 2000. Next, the number of minutes per week spent in light and moderate to high intensive activities is calculated via an MET threshold. We mentioned in the introduction that this estimate is inaccurate and often biased. If we succeed to find a better estimate for intensity, then we might use the SQUASH questionnaire to quantify the size of the bias.

2.2 Preprocessing of measurements

The sensor data contains a timestamp. Each sensor is synchronised using the time stamp variable, the measurements of the other sensors and the beginning and end of the lab activities. The sensors are synchronised with the logbook of activity by visual inspection of one variable

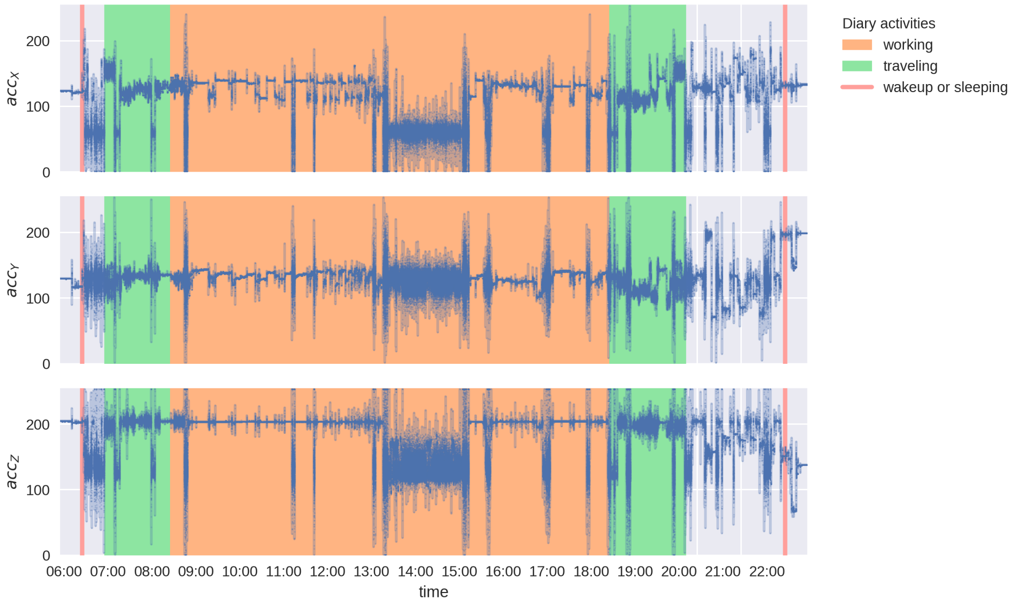


Figure 2.3 Accelerometer data of one respondent during one day.

from each sensor. We corrected the timestamps, such that the jumps occurred at the right time and the other activities looked fine. Two files with time shifts at the beginning and/or end were produced: time shifts of the activities and time shifts of the sensor. 19.85 % of the sensors and 19.49 % of the activities contain a time shift, which stresses the importance of data cleaning. During the data cleaning, we encountered the following situations:

- Vyntus contained some duplicated time series
- Start- end/or end times of activities do not precisely correspond to logbook
- Some sensor output was damaged, and could not be cleaned
- There is not always enough time between activities for the heart rate to calm down.

An accelerometer at rest on the surface of the Earth will measure an acceleration due to Earth's gravity, straight upwards (by definition) of $g = 9.81$ m/s2. By contrast, accelerometers in free fall (falling toward the centre of the Earth at a rate of about 9.81 m/s2) will measure $g = 0$ zero. However, it turns out that most sensors are not calibrated, so that they often measure nonzero acceleration in all directions. Algorithms exist to remedy this at the end of the data collection (e.g., van Hees et al., 2014), but we were not aware of this at the time of analysis.

Variables like heart rate and VO_2 need a warming up period. The length might differ per activity, e.g. cycling has a longer warming up period than for example walking. The variation in accelerometer data depends on the activity. Since UKK is averaged per second, the time series are quite smooth. The five jumps are clearly visible in Hexoskin and ActivPAL accelerometers, but not in UKK data. Hence, a frequency of one second is likely too rough to detect activities.

Compared to the logbook, the quality of the diaries was quite bad. Most respondents will have filled in the diary at the end of the day, with the result that begin and end time of activities would be guessed. One respondent did not have weekdata and three other respondents did not keep a diary. We compared the diaries to ActivPAL MET values and number of events in order to clean the diaries and assess their quality. During the data cleaning we identified two diaries of insufficient quality. The following observations are made during the data cleaning:

- Only 75% percent of the respondents wore the sensor for at least 6 and a half days, while respondents where instructed to wear the sensor for 7 days.
- The begin- and endtimes of activites are often not accurately.
- Some diary items are probably wrong, but we were unable to correct the timestamps.
- The diary contains obvious mistakes, for example incorrect dates and activities without an endtime.

Table 2.2 shows the number and duration of activities. Note that the standard deviation is quite large for all activities, meaning that the duration differs between respondents and over the days. Respondents typically cycle for a short period. The maximum duration of sports is 6 hours, and 90% quantile is 2 hours. If we compare the median duration with the margin of 15 minutes, the quality of the diary is not very good. Therefore, we will only use the diary in combination with (ActivPAL) predictions.

Table 2.2 Number of times activities occur in diaries, median and standard deviation of duration (in minutes).

	# respondents	# activities	duration (std.)
Non-wear	28	104	20 (78)
Cycling	31	314	10 (18)
Traveling	33	472	30 (130)
Sports	24	69	60 (53)
Working	30	123	420 (197)

3 Intensity

Estimating the intensity of physical activity is a crucial element if one wants to determine if people adhere to physical activity guidelines. This section describes three metrics, MET, heart rate and tri-axial Medium Amplitude Deviation (MAD), that are related to the intensity of physical activities. Moreover, we explore the use of thresholds to classify intensities into three classes. The use of thresholds, coupled with counts of the number of times an acceleration exceeds a certain threshold has long been the only way to determine the intensity of movement. With the onset of machine learning, the focus has shifted to recognising activities and subsequently allocating MET values to those activities.

3.1 Metabolic Equivalent of Task

One MET corresponds to the intensity of quietly sitting and is equal to approximately 3.5mL VO_2/kg/minute. The Vyntus sensor measures oxygen intake. By combining measurements of Vyntus with the weight of respondents, we can compute a reliable estimate for energy expenditure from Vyntus measurements, expressed as the MET value.

ActivPAL's software provides MET values. The software assigns a MET value of 1.25 to the activities non-wear, (primary/secondary) lying, sedentary, travelling. Standing has a MET value of 1.4. The other activities, stepping and cycling, have a MET value between 1.8 and 5.1. The empirical distribution of stepping MET values during the laboratory sessions contains two peaks. The distribution of cycling does not contain a clear peak, indicating that ActivPAL is unable to separate cycling light and heavy in the operationalisation chosen in the lab setting, where cycling occurred at 60 cycles per minute for each intensity level.

We can compare activPAL's MET value with the baseline from Vyntus for the laboratory activities. We calculated the MET values for respondents for which both ActivPAL and Vyntus data are available. The measurements might need some time the converge. Therefore, we removed a warming-up period at the start of each activity. The length of the warming-up period depends on the type of activity. The MET values are averaged per respondent and activity over 10-second intervals by computing the mean value. Vyntus VO_2 corrected for the weight of respondents correlates well with activPAL's MET. We added the Pearson's correlation for each respondent and then averaged over the number of respondents. This resulted in an average correlation of 0.836 over the time series of 10-second average MET values with a standard deviation of 0.081. Even thought the average correlation is quite high, there can still be a large absolute difference between the MET of Vyntus and ActivPAL for certain activities and respondents.

Figure 3.1 shows the MET value of both sensors. As mentioned, the activities sitting and standing have a constant MET value from ActivPAL. The median MET by the Vyntus sensor suggests that there are different MET values possible for these low-intensity activities. However, by using Vyntus MET it is hard to distinguish the activities sitting and standing. The MET values of sitting and standing are quite different from the other activities for both sensors. The highest MET values occur for activity running for both sensors, although there is much variation in the MET values of running. The highest observed MET value from Vyntus is 11.4 versus a maximum of 5.1 of activPAL's MET. For high-intensity activities, the MET values of Vyntus are more volatile, especially after reducing the sampling frequency to 5 seconds or less. On the other hand, ActivPAL possibly suffers from classification mistakes, since the MET value depends on the activity. In addition, MET values for the ActivPAL are censored at 5.1 MET as a result of the sampling resolution and range.

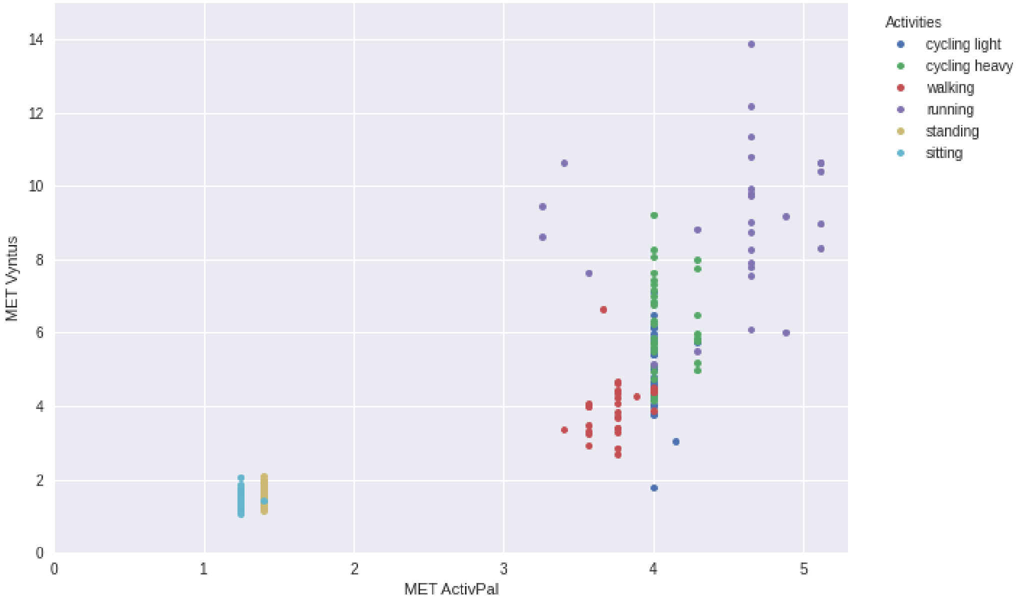


Figure 3.1 Median MET from ActivPAL per respondent and activity versus median MET as calculated by Vyntus VO_2.

3.2 Heart rate

It is possible to measure energy expenditure accurately when using a combination of an accelerometer and heart rate (Brage et al., 2004). Therefore, we are firstly interested in the correlation between heart rate as measured by Hexoskin and MET. Figure 3.2 shows that a higher value of heart rate corresponds in general with a higher value of VO_2 after correcting for body weight (in kg) and vice versa. The value of VO_2 differs between activities, but the VO_2 of low-intensity activities sitting and standing is often very similar.

Pearson's correlation per respondent of VO_2 and heart rate after resampling to a frequency of 10 second intervals range over the respondents from 0.53 to 0.98 after removing a warm-up time. The average correlation over the respondents equals 0.896 with a standard deviation of 0.084. Low correlation is often caused by a relatively high heart rate of single activity. Possibly, some respondents have performed high intensity movements just before the activity started, so that the heart rate did not have time to return to the baseline heart rate. The baseline heart rate varies over respondents as a result of age and fitness variation. It will be interesting to compare the difference between current and baseline heart rate with VO_2 values in future research. In Section 2 we will combine Hexoskin heart rate with accelerometer data to examine the explanatory value of heart rate.

3.3 Mean/median amplitude deviation

The mean amplitude deviation is the measure of intensity used by the UKK software. It has shown to perform well to accurately separate intensity of activity compared to other statistics like the skewness, kurtosis and difference between high and low percentiles (Vähä-Ypyä, Vasankari, Husu, Suni, et al., 2015). While UKK uses mean deviation for their calculations, we

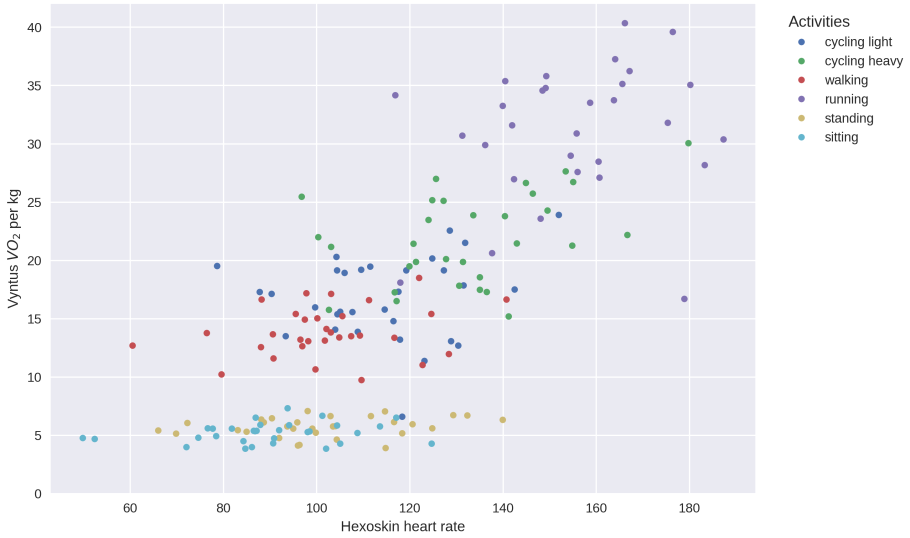


Figure 3.2 Median heart rate versus median VO$_2$ per respondent.

chose to use the median instead. Some of the activities showed large within variations that would unduly influence the amplitude deviations.

The Median Amplitude Deviation (MAD) is calculated per axis r from raw accelerometer data. The MAD is the median of the deviations from the median acceleration during a short epoch, for example a few seconds. $MADx_t$ is the MAD of the x-axis (direction) during epoch t:

$$MADx_t = \text{median}\left(|x_{t,1} - \hat{x}_t|, |x_{t,2} - \hat{x}_t|, |x_{t,3} - \hat{x}_t|, \ldots, |x_{t,k} - \hat{x}_t|\right) \tag{3.1}$$

where $\hat{x}_t = \text{median}\left(x_{t,1}, x_{t,2}, x_{t,3}, \ldots, x_{t,k}\right)$ is the median over the epoch and k is the number of observations in epoch t. The $MADy_t$ and $MADz_t$ can be calculated based on respectively the y-axis and z-axis accelerometer values. The three axes can be summarized in the $MADxyz_t$.

$$MADxyz_t = \text{median}\left(|r_{t,1} - \hat{r}_t|, |r_{t,2} - \hat{r}_t|, |r_{t,3} - \hat{r}_t|, \ldots, |r_{t,k} - \hat{r}_t|\right) \tag{3.2}$$

The variable $r_{t,i} = \sqrt{x_{t,i}^2 + y_{t,i}^2 + z_{t,i}^2}$ for i in $1, \ldots, k$ summarises the information from the three axis. The epoch length k is a parameter of this measure of intensity. The provided mean amplitude deviation for UKK contains 6-second epochs (Vähä-Ypyä et al., 2018). Another parameter is the step size that determines the frequency of the MAD. A step size equal to the epoch length means non-overlapping epochs.

Table 3.1 shows the correlation between the MAD values calculated for ActivPAL, UKK, and both IMUs with VO$_2$. The mean amplitude deviation of UKK is provided with a 6-second epoch length and a step size of 1 second. The time series of all sensors are down-sampled to 6-second intervals for a fair comparison between the sensors. We calculated the correlation between activPAL's MAD and and VO$_2$ for different epoch lengths, to get a first impression of the effect of the epoch length on the correlation. For an epoch length between 5 and 25 seconds, the correlation of activPAL's MAD and VO$_2$ varies between 0.791 and 0.795. The correlation is quite robust for epoch length if the epoch length is between 5 and 25, but decreases for shorter or longer epoch lengths. Note that the optimal epoch length might also depend on the frequency of the sensor. The accelerometer values are not re-scaled before calculating the MAD. As a result, the range of MAD values differs per sensor. The MAD of UKK varies between 4 to 1233, while the MAD of ActivPAL varies between 0 and 47.512. Moreover, note that UKK's $MADxyz$ is calculated using the mean instead of the median. The median in Equation 3.1 is replaced with a mean and

$\hat{r}_t$ is also calculated by taking a mean (see Equation 3-6 in Vähä-Ypyä et al., 2018). Unfortunately, we cannot calculate the MAD of UKK by using the median, since the received UKK data are already aggregated.

Table 3.1 shows that the correlation of UKK's mean amplitude deviation and VO_2 is lower than the correlation of ActivPAL. However, it is not straightforward to compare the correlations, since they are calculated by a different method (median versus mean) and based on a different set of respondent activities.

Table 3.1 Pearson's correlation between MAD with 6-second epochs and Vyntus VO_2.

	mean correlation	standard deviation	# respondents
ActivPAL	0.791	0.089	29
UKK	0.630	0.164	30*

* 13 respondents lack data for activity sitting, 2 respondents lack data of activity standing

The MAD of ActivPAL has the highest correlation with Vyntus VO_2, therefore we compared the mean MAD of ActivPAL per respondent and activity versus Vyntus VO_2 corrected for body weight. The MAD in Figure 3.3 shows three groups of clustered activities: standing and sitting, cycling light and heavy with walking and running. However, it is difficult to distinguish all activities. The MAD value of cycling light and heavy is very similar, while the VO_2 per kg differs.

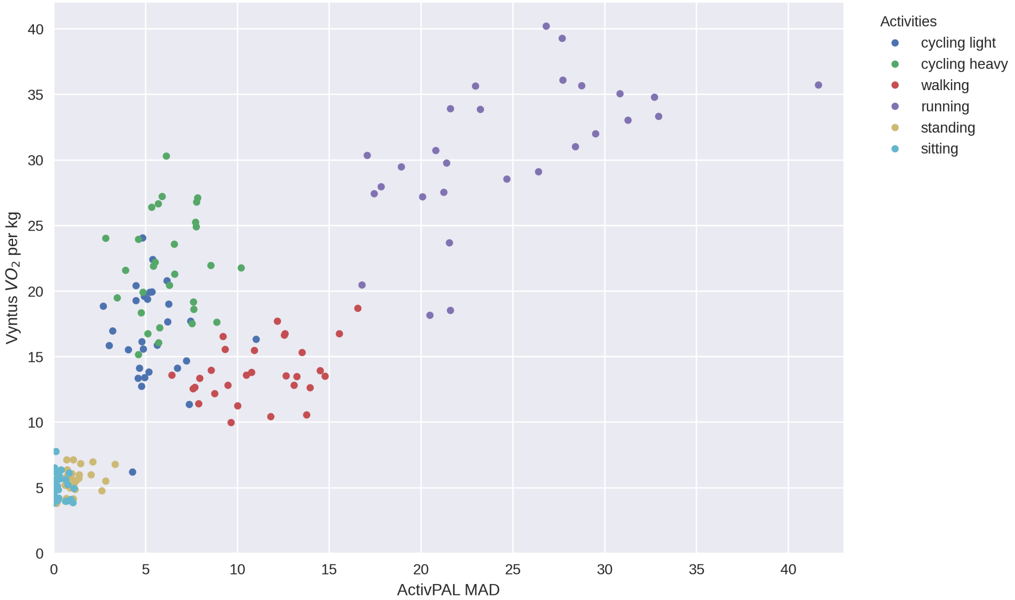


Figure 3.3 Median MAD versus median VO_2 per respondent after subtracting a warming up period of at most two minutes.

3.4 Three intensity classes

Recall that we are ultimately interested in the time spent in moderate to high intensity activities. The threshold between low and moderate activities is 3 MET. Using specific cut-off points or thresholds, one can determine if the intensity is low, moderate of high. Intensity thresholds can vary over populations (e.g., Kuppevelt et al., 2019). In this paragraph we present an algorithm to compute the degree of intensity directly from the raw accelerometer data of the ActivPAL device. A similar rule-based approach can be found in Staudenmayer et al., 2015. The degree of intensity is computed per minute and can take the values: low, moderate or high. The general idea is that when during a certain period the changes of the accelerometer values are below or above given thresholds, the intensity of activity is low, moderate or high.

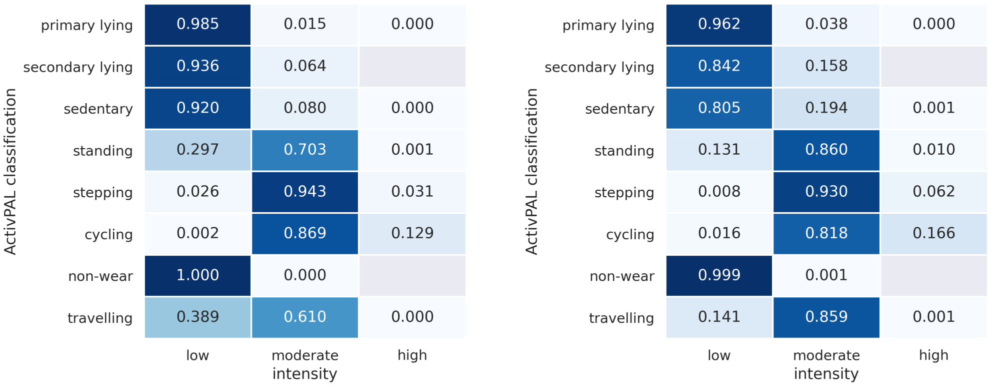
The pseudo code is shown in Algorithm 1. First, for every second the median of the accelerometer values is computed. Next, we compute for every second the MAD for a rolling window of size w seconds and step size 1 second. As a result we get for every second the MAD values of the x-, y- and z-axis of the accelerometer. The next step is to count the number of seconds these MAD values are below a threshold t_s^l within one minute. If this count is above a threshold t_m^l the degree of intensity is low. If this count is below this threshold, first compute the geometric mean of the MAD values of the axes ($MADxyz_i$). Next, count the number of seconds this value is above a threshold t_s^h within one minute. If this number is above a threshold t_m^h the degree of intensity is high, otherwise the degree is moderate.

Algorithm 1 Compute intensity levels

```
Input: Time series $X, Y, Z$ with integer accelerometer measurements, epoch length $k$, thresholds $[\tau_m^l, \tau_m^h]$
    and $[\tau_s^l, \tau_s^h]$ per minute ($m$) or per second ($s$).
Output: Intensities $M_t$ for all minutes $t$
for every second $t$ do
    Calculate $MADx_t$, $MADy_t$, $MADz_t$ and $MADxyz_t$ from $X, Y$ and $Z$ using Equation 3.1 and 3.2
    if $MADx_t < \tau_s^l$ and $MADy_t < \tau_s^l$ and $MADz_t < \tau_s^l$ then
        $S_t^l \leftarrow 1$
    else
        $S_t^l \leftarrow 0$
    end if
    if $MADxyz_t > \tau_s^u$ then
        $S_t^h \leftarrow 1$
    else
        $S_t^h \leftarrow 0$
    end if
end for
for every minute $t$ do
    $c_t^l \leftarrow \sum_{i \text{ in minute } t} S_i^l$
    $c_t^h \leftarrow \sum_{i \text{ in minute } t} S_i^h$
    if $c_t^l > \tau_m^l$ then
        $M_t \leftarrow$ low
    else if $c_t^h > \tau_m^h$ then
        $M_t \leftarrow$ high
    else
        $M_t \leftarrow$ medium
    end if
end for
```

Figure 3.4a shows the result of applying this algorithm for one set of thresholds. The activities are those computed by activPAL's software. The values of the thresholds are $\tau_m^l = 40, \tau_m^h = 50, \tau_s^l = 1, \tau_s^h = 6$ and the epoch length is 6 seconds with a step size of 1 second, yielding 60 MAD values per minute. The intensity values are calculated for both labdata and weekdata of 33 respondents. The thresholds are derived by comparing the computed degree of intensity with the activities of one participant during the lab session. Additional research is necessary to find thresholds that generalise well. Thresholds can also be dependent on demographics like sex, age, lifestyle etc. See Sasaki et al., 2011; Troiano et al., 2008; Freedson et al., 1998; and Matthews, 2005 for a discussion of thresholds for specific devices and specific positions on the body.

As can be seen in Figure 3.4, low MAD intensity coincides quite well with the ActivPAL classifications of primary and secondary lying, sedentary behaviour and non-wear. Stepping and



	low	moderate intensity	high
primary lying	0.985	0.015	0.000
secondary lying	0.936	0.064	
sedentary	0.920	0.080	0.000
standing	0.297	0.703	0.001
stepping	0.026	0.943	0.031
cycling	0.002	0.869	0.129
non-wear	1.000	0.000	
travelling	0.389	0.610	0.000

(a) The median is used to compute the amplitude deviation.

	low	moderate intensity	high
primary lying	0.962	0.038	0.000
secondary lying	0.842	0.158	
sedentary	0.805	0.194	0.001
standing	0.131	0.860	0.010
stepping	0.008	0.930	0.062
cycling	0.016	0.818	0.166
non-wear	0.999	0.001	
travelling	0.141	0.859	0.001

(b) The mean is used to compute the amplitude deviation.

Figure 3.4 ActivPAL's activity classification versus intensity as calculated by Algorithm 1 based on the lab- and weekdata.

cycling are most often classified as moderate activity, although cycling is also quite often classified as high intensity. There are two categories with low agreement: both standing and travelling are classified as moderately intensive in MAD. Comparison with diaries has determined that ActivPAL not always correctly determines travelling. Car and train travel are mostly correctly recognized as travel, but travelling by bus and tram are not. The MAD calculation seems to suffer from the same over-estimation of activity. Although some activity during travel is probable (walking to the train door), most of the movement measured is more probably the result of bumps in the road.

Our MAD is calculated by taking the *median*, while UKK prefers a *mean* to compute the amplitude deviation. The choice for the median is based on two arguments. Firstly, the median is more robust to large deviations. Low-intensity activities such as sitting contain strong deviations during a short period. We expect that these deviations occur for example when a respondent changes the position of his legs. These deviations shouldn't have a large effect on the intensity. Secondly, the performance of the MAD with using a median is overall better than the mean. Figure 3.4b shows the confusion matrix of the intensity categories compared to activPAL's classification. The low physical activities primary and secondary lying, sedentary behaviour are more often classified as low-intensity activity when the median is used instead of the mean. The two confusion matrices differ substantially for the activities standing and travelling. Ideally standing is classified as a low-intensity activity. Hence, the amplitude deviation based on the median performs better for standing. Travelling is a very generic category. Travelling by public transport can be classified as low intensity, for example, if respondents sit in a bus. Travelling by bicycle can also be classified as more intense.

3.5 Short summary of findings on measuring intensity

In this Section we compared three measures of intensity, MET, heart rate and MAD (with thresholds) with Vyntus VO_2 to answer the first research question: How can we estimate intensity based on the lab- and weekdata? The MET value that is calculated by ActivPAL can be used to distinguish standing and sitting from the more intensive activities, but the MET values of the activities running, cycling, walking do not cluster well. Hence, we conclude that activPAL's MET value performs moderately well. The variation of Hexoskin's heart rate over the respondents is quite high. However, the heart rate is able to distinguish activities for a given respondent. Heart rate could be a promising measure of intensity, provided that inter-person variation in baseline heart rate is corrected for. The MAD values calculated on the ActivPAL data are able to distinguish three groups of activities: standing and sitting, walking and cycling and running. A threshold approach classified the MAD values into low, moderate and high intensity classes. We conclude that the MAD is a good estimate of intensity, as it performs better than activPAL's MET and Hexoskin's heart rate. However, the thresholds need further refinement.

4 Activity classification

The goal of time series classification is to identify each (group of consecutive) observations as coming from one of the predefined classes. Time series classification differs from other classification tasks like image classification since it involves a temporal aspect. Because of the temporal aspect, it is known as a challenging problem in data science. Machine learning is one of the most used approaches to classify a time series. There are mainly two approaches within machine learning: unsupervised learning like clustering algorithms and supervised learning which requires a training set with labels that is used to train the mode how to differentiate between classes. Human activity classification is the general term for the task of classifying measurements into classes of physical activities. Our problem involves the classification of the time series that are measured by the sensors. Since the true activity is known during the laboratory sessions, supervised algorithms will be applied. Since there are more than two activities, the task is known as multi-class classification.

We do not know if it is possible to differentiate 100 % accurately between all activities using the measurements from the sensors. On the one hand, two identical time series may belong to different activities. On the other hand, very different time series might be generated during the same activity due to respondent heterogeneity. Therefore, we cannot expect an accuracy of 100 %. We will compare the accuracy of different types of algorithms, to get an impression of the accuracy that can be achieved.

4.1 Train, validation and test set

Our machine learning model is trained by providing features, the input variables, and the true activity, the labels. The output of the model is compared with the labels to evaluate the performance and update the model parameters. The features are X, Y, and Z-acceleration data (possibly supplemented with heart rate) at a regular sample frequency. The data are normalised for each feature by subtracting the mean and dividing by the standard deviation. The labels are the classes of activities at the same frequency of the features. There are only labels for the laboratory sessions.

The steps from raw input data to features are called feature extraction. Firstly, the accelerometer data are re-sampled at a frequency of 10 seconds. Secondly, the data are divided into short periods that are classed slices. Each slice of a fixed length k corresponds to at most one respondent and one activity. If the number of observations after re-sampling is not a multiple of k, then a few observations might be disregarded. Slice s_t contains the three-dimensional time series from timestamp $t-k$ until t

$$s_t = \left\{ \left\{X_j\right\}_{j=t-k}^{t}, \left\{Y_j\right\}_{j=t-k}^{t}, \left\{Z_j\right\}_{j=t-k}^{t} \right\},$$

for $k \geq j \geq n$. For a given respondent and activity, the training data equals of the time series $[s_0, s_m, s_{2m}, \ldots, s_{n-k}]$ for a given $1 \leq m \leq k$. We have encountered k in Section 3 as the epoch length and m as the step size. A small value of m corresponds to highly overlapping slices, while there is no overlap for $m = k$. The slices of activities are sorted by timestamp for each respondent. The label $y_{t,a}$ of a slice is the activity of the respondent of interest at the last timestamp of the slice.

$$y_{t,a} = \begin{cases} 1, & \text{if activity } a \text{ occurred for at time } t \\ 0, & \text{otherwise} \end{cases}$$

The features and labels are divided into a train, validation, and test set by sampling respondents uniformly at random. Firstly, the respondents are divided into a train and test respondents. The train (test) set consists of all the measurements of the train (test) respondents. Since all respondents performed the same activities during the laboratory sessions, the number of time slices of each respondent is roughly the same. Secondly, the train set is divided into an actual train set and validation set. 5-fold cross-validation is used to estimate hyper-parameters; the training data are divided into five different splits of 80% train data and 20% validation data.

With 40 respondents and cleaned data of only 36 respondents, it might matter which respondents are in the test and training set. In other words, the model performance on the test respondents might be dependent on the train-test split. Therefore, we repeated the model training and evaluation for 5 different splits in train/validation and test sets and averaged the results.

4.2 Neural networks

Two types of machine learning models are used to classify the time series: Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. These types of neural networks are known to work well for time series classification. A feed-forward neural network describes a single differential score function that maps the input data to an output vector. The network can be visualised as a directed graph that consists of nodes and edges. The nodes, also called neurons, are ordered in layers and contain functions that are applied to the data. The edges describe the data flow from the input layer to the output layer. The fully connected neural network is one of the simplest types of networks. In a fully connected network, each node in a layer directs to each node in a consecutive layer. With only a few hidden layers, the number of edges in a small network is already quite high. The output layer is always fully connected.

The shape of the input data is equal to the number of features times the input shape k. If the features are X, Y, and Z-acceleration data, then the number of features equals 3. The output vector $\hat{y}_{t,a}$ has equally many elements as the label $\hat{y}_{t,a}$. The number of nodes of the hidden layer depends on the type of layer and/or the number of coefficients. The number of layers, size of layers, and type of layers can be referred to as network architecture.

A linear function $f(x) = Wx + b$ is applied to the input of each node. The parameters weight W and bias b will be estimated during the training phase. An activation function $\sigma(\cdot)$ is applied to the result of the matrix multiplication. We use the well-known rectified linear activation function (ReLU) of the form $\sigma(f(x)) = \max\{0, f(x)\}$. Note that ReLU function is not continuous at 0. The final layer contains a softmax activation function, to map the values to a vector. The i^{th} component of the output vector equals

$$\sigma_{\text{softmax}}(x)_{t,i} = \frac{e^{x_{t,i}}}{\sum_{j=0}^{a} e^{x_{t,j}}}$$

such that each value is between 0 and 1 and $\sum_{j=0}^{a} e_{x_j} = 1$. A common approach is to use the argmax as the predicted class.

The loss is minimised during the training phase. Each training iteration contains three steps. Firstly, the loss is computed for a batch of observations. Secondly, the gradient of the weights is determined via back-progagation. Thirdly, the learnable parameters are adjusted according to the gradient. The size of the adjustments is described by the learning rate. The biases are initialised with zeros and weights with samples from a normal distribution with mean 0 and standard deviation 0.01 to avoid getting stuck at an all-zero weight solution.

The iterative procedure is repeated until a stopping criterion is met. The easiest stopping criterion is to stop when all observations are evaluated. However, the network might benefit from seeing observations more than once. The number of times the complete training set is evaluated is called the number of epochs. A second option for a stopping criterion is, therefore, to stop after a fixed number of epochs. However, the model might over-fit the training data if the number of epochs is too large. Note that the number of learnable parameters is already large for a single-layer hidden network, so over-fitting can occur easily.

There are two common approaches to prevent over-fitting: regularisation and a refined stopping criterion. Our models contain a dropout layer in which nodes are removed at random during the training phase. As a result, the network will be more robust and is less likely to over-fit the train data. Another common approach is to add a regularisation term to the loss function. Next to a dropout layer, we refined the stopping criterion. The training is stopped when the loss does not decrease during two epochs, and the best model in terms of validation loss is returned.

4.3 Sensor comparison

Recall that we are interested in the performance of machine learning models trained on labdata in order to answer the first research question. To this end, a different CNN model if trained for each sensor. The features are based on X-, Y-, Z-acceleration data of four sensors, ActivPAL, IMU wrist and tibia, and Hexoskin. Heart rate measured by Hexoskin is an additional feature. Note that the sensors are placed on different positions on the body, and therefore we do not only compare the technology to measure acceleration, but also different positions. Raw UKK accelerometer data are not available, therefore we have not estimated machine learning models on UKK data.

In general, the performance of a trained machine learning model depends on the size of the training data. The number of respondents for IMU worn on tibia and wrist are respectively 18 and 12, which might affect the model performance. One respondent does not have Hexoskin data including heart rate. The test respondents of ActivPAL are chosen uniformly at random. To compare the model performance of the different sensors, the test respondents of Hexoskin and IMU are a subset of the ActivPAL test respondents. The remaining respondents are used for training.

Figure 4.1 shows the layers of the network. The input data are divided into batches of 16 observations. Each convolutional layer is followed by a pooling layer to reduce the spatial size of the representation. Including a pooling layer between convolutional layers is a common practice. The first convolutional layer has shape 64 and the second layer shape 128. ReLU is used as an activation function. The dropout layer contains a dropout rate of 0.5. The optimiser is RSMProp and the learning rate equals $1e - 4$. Accuracy and balanced accuracy are used to evaluate the performance of the models.

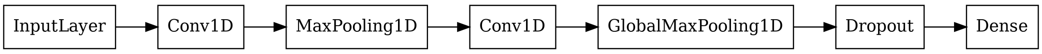


Figure 4.1 Layers of CNN.

We will measure the performance of the models in accuracy and balanced accuracy. Both metrics can be derived from the confusion matrix. The confusion matrix divides classifications into true positives (TP; observations correctly classified as belonging to a certain activity), true negatives (TN; observations correctly classified as not belonging to a certain activity), false positives (PF; observations incorrectly classified as belonging to a certain activity), and false negatives (FN; oservations incorrectly classified as not belonging to a certain activity). The accuracy can be calculated as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Balanced accuracy weights all activity classes equally. Other frequently used metrics include the precision and recall that can also be calculated from the confusion matrix and the specificity and sensitivity. The recall and precision focus on the true positives, whereas the sensitivity and specificity focus on the proportion of correctly classified observations. These metrics can be calculated for the train, validation and test set.

The performance of models with different step sizes and train percentages are shown in Table 4.1. The validation accuracy is always higher than the test accuracy, since the parameters are optimised using the validation set. Models with the feature heart rate perform better for all sensors. Some models are trained on a smaller training set since the number of respondents with cleaned data varies over the sensors. The lower performance of IMU wrist is probably due to the low number of training respondents. The model trained on ActivPAL acceleration data performs best. The step size and training percentage are optimised by comparing the validation accuracy. A step size of 25% of the epoch length yields better results than 50%, and the training percentage of 0.6 is slightly better than 0.7 for all sensors except ActivPAL.

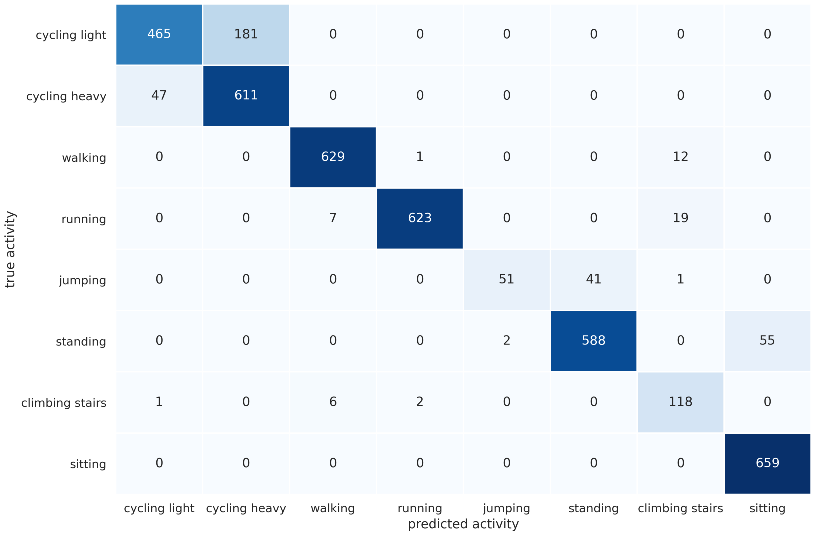
Table 4.1 (Balanced) validation and test accuracy of the best model for each sensor.

sensor	heart rate	accuracy	balanced accuracy	validation accuracy
Hexoskin	True	0.751	0.746	0.976
	False	0.724	0.722	0.955
ActivPAL	True	0.909	0.872	0.975
	False	0.875	0.846	0.965
IMU tibia	True	0.879	0.861	0.991
	False	0.730	0.733	0.982
IMU wrist	True	0.668	0.647	0.975
	False	0.449	0.416	0.946

Figure 4.2 shows the confusion matrix of the best sensor: ActivPAL including the heart rate. Cycling is performed on a home trainer, so cycling heavy instead of light means shifting to a higher gear which might be hard to detect by a sensor on the upper thigh. Cycling light and heavy are indeed often confused. Note that cycling is not confused with other activities, such as walking and standing. As a result, the sensitivity and specificity of cycling compared to the other activities are both quite high. Using the confusion matrix, we can calculate the plausibility that a slice of data that is classified as a certain activity is correctly classified: the positive likelihood ratio. Cycling light has sensitivity of 0.72 and a specificity of 0.99 compared to all other activities, which results in a positive likelihood ratio of 52.08. Cycling heavy on the other hand as a sensitivity of 0.93 and a specificity of 0.95 compared to the other activities, which results in a positive likelihood ratio of 17.76. These likelihood ratios show that the classification is able to distinguish cycling light and heavy, although improvements are possible.

Jumping are less accurately classified, perhaps because this activities was performed during a shorter period. Jumping has a sensitivity of only 0.55. However, the specificity is very high. Respondents stood still between the five jumps, hence the model might confuse jumping with standing. The model might confuse sitting with standing since these are both low intensity activities. It should be noted that some sensors were not properly calibrated before the laboratory sessions, this might (partly) cause the confusion between standing and sitting. Figure 4.2 shows that standing and walking are well classified.

Section 3 described the correlation between heart rate and VO_2. On average, the heart rate differed per activity. Especially when the accelerometer data has trouble to distinguish activities like cycling light and heavy, we hope that heart rate can improve the classification. After using five different train-test splits to train a model with and without heart rate, we concluded that heart rate improved the classification on average. However, all categories are improved a little on average, not just cycling. However, heart rate needs a warming-up (or cooling-down) period when changing from a low to heavy intense activity (or vice versa).



true activity \ predicted activity	cycling light	cycling heavy	walking	running	jumping	standing	climbing stairs	sitting
cycling light	465	181	0	0	0	0	0	0
cycling heavy	47	611	0	0	0	0	0	0
walking	0	0	629	1	0	0	12	0
running	0	0	7	623	0	0	19	0
jumping	0	0	0	0	51	41	1	0
standing	0	0	0	0	2	588	0	55
climbing stairs	1	0	6	2	0	0	118	0
sitting	0	0	0	0	0	0	0	659

Figure 4.2 Confusion matrix for ActivPAL with heart rate. The numbers correspond to the number of observations in the test data (without overlapping slices).

The confusion matrix of activPAL's predictions versus the model predictions is shown in Figure 4.3. This confusion matrix is based on a frequency of one minute and only the labdata. The activPAL software uses a different set of possible activity classes: non-wear, standing, running, cycling, stepping, primary lying, secondary lying, travelling and sedentary. Primary lying occurs mainly during the night, while secondary lying occurs during the day. The classification can distinguish for example traveling but does not have a specific class for e.g., climbing stairs. None of the observations during the laboratory sessions is classified as non-wear or traveling as expected. The software predicts only four distinct classes during the laboratory sessions. Both activity taxonomies contain for example the class cycling. There are some discrepancies between the true status and the activPAL prediction. This is partly the result of the fact that not all measured activites are classified in activPAL. For example, climbing stairs is mostly classified as stepping, sometimes also as cycling. However, on the classe in common, cycling was classified correctly in 94 % of cases, stepping in 90 %, sitting in 94 %, and standing in 93 % of cases. Discrepancies may be the result of of differences with the training data from activPAL's model, but also of slightly blurry edges between activities and slight errors in timing of activities.

4.4 Model comparison

In this subsection we will compare different models based on activPAL's accelerometer data. The goal of these model comparisons is to get an impression of the possibility to increase the

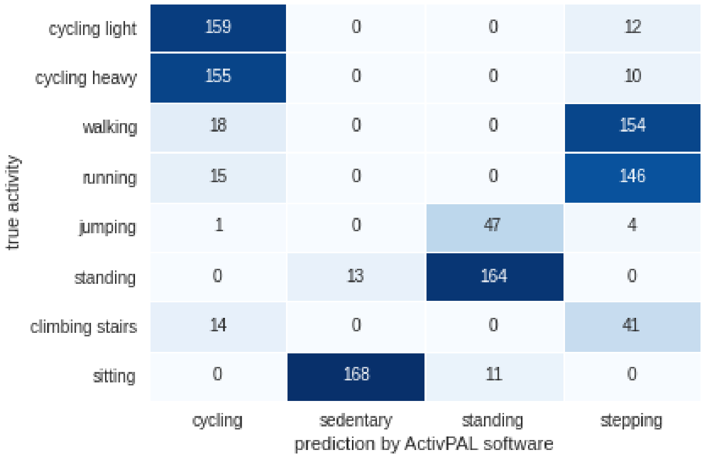


Figure 4.3 Confusion matrix of ActivPAL predictions during laboratory sessions. The confusion matrix shows cases with at least one prediction.

accuracy by changing model type and finetuning hyperparameters. The current model contains a lot of learnable parameters. While the task of classifying activities is quite difficult, the number of parameters might be too high. If a smaller model is able to classify the activities equally well, we prefer the smaller model that needs probably also less epochs for training.

The model trained on ActivPAL accelerometer data without heart rate will be optimized by tuning hyperparameters. Similarly to the models from Table 4.1 the input shape is 16 as batch size, 3 for all directions of acceleration and 100 as chosen slice length. The validation loss is minimized, while the model keeps track of the (balanced) accuracy in order to evaluate the performance. We performed grid search with only two or three values per hyperparameter. The resulting model can likely be finetuned into a slightly better model, but the hyperparameter optimization gives insight in the increase in accuracy. The following values of hyperparameters result in 288 different combinations in the grid search:

- Activation function: ReLU or ELU.
- Batch size: series of 16 or 64 observations.
- Dropout rate: 0.4, 0.5 or 0.6. A higher dropout rate might reduce the risk of overfitting.
- Number of epochs: 5, 10 or 20. Experiments indicated that increasing that training is often stopped before the 20 epochs of training have finished because of the early stopping criterium.
- Number of convolutional layers: two or three, each layer is followed by a pooling layer.
- Number of neurons. The number of neurons is either increasing per layer ($[64, 128, 258]$ for a 3-layer network) or the same for all layers ($[64, 64, 64]$). A larger model is able to detect more patterns, but can also overfit the data.
- Optimizer. Two commonly used optimizers are used: Adam optimizer and RSMprop.

Table 4.2 contains the (balanced) accuracy of the three best models for activPAL without heart rate. The table also contains a standard deviation based on five different splits of train and test data. The test set contains no overlapping slices. Over the 288 different combinations of hyperparameters, the accuracy is always larger than 0.8, meaning that a smaller model with 2 hidden layers is also able to classify the activities accurately. The best models are trained for maximal 20 epochs, which is an increase compared to the default model with 10 epochs. The number of neurons increase per layer. The minimal balanced accuracy of the 288 models is 0.789, while the maximal value of 0.818.

Table 4.2 Parameters and accuracy of three models with highest validation accuracy in the CNN hyperparameter optimization.

Rank of mean val. accuracy	1	2	3
Activation function	Relu	Relu	Relu
Batch size	16	16	64
Dropout rate	0.4	0.4	0.6
# epochs	20	20	20
# conv. layers	2	2	3
# neurons	[64, 128, 256]	[64, 128, 256]	[64, 128, 256]
Optimizer	RSMprop	Adam	Adam
Mean accuracy	0.83	0.83	0.823
Std. accuracy	0.032	0.032	0.038
Mean balanced accuracy	0.808	0.808	0.801
Std. balanced accuracy	0.028	0.028	0.032

Figure 4.2 shows that it is difficult to differentiate between cycling light and heavy. The same holds for the best model of the hyperparameter tuning. Therefore, we will combine the two actvities into one activity called cycling from here on. Using a two-step classification we might first classify the activity as cycling and afterwards try to differentiate between cycling light and heavy. However, the distinction between cycling light and heavy is not very important for now.

The confusion matrices of the 5 different train-test splits for the best model are shown in Appendix A Figure A.1 until A.5. Note that the difference in accuracy between models with different test sets is quite high. This might be due to the relatively low number of respondents and/or respondent hetergeneity.

We benchmark the performance of the optimized CNN with a Long-Short Term Memory (LSTM) model and a Random Forest (RF) classifier. LSTM is a type of neural network that is often used for time-series classification since it can remember observations over time intervals. This might be useful to classify activities that consist of multiple (repeating) movements such as upward and downward movements during cycling. RF is a classical machine learning algorithm that combined multiple decision trees to classify activities. The algorithm uses less parameters, therefore it is interesting to compare the difference in performance.

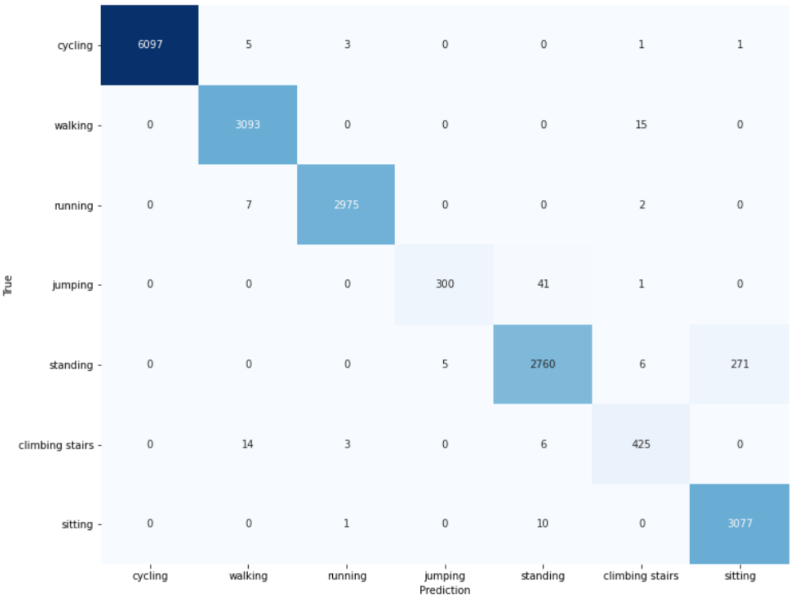
Random Forest requires one-dimensional feautres, therefore we need to transform the X-, Y-, and Z accelerometer time series. The distribution of frequencies during a slice is computed using Fast Fourier Transformation (FFT). The 20 peak frequencies in the slice were used as the features of the model. The slice length is set to 20 seconds while calculating the features, with step size m of 1 second. An illustration of a spectogram of a respondent during an activity is shown in Appendix A figure A.6. Next to the peak frequencies, average accelerometer data for the three axes are used as features. LSTM is implemented in tensorflow, with the Scikit-learn implementation of RF classifier[3]. RF classifier contains 100 estimators. Default values are used for other hyperparameters of the RF and LSTM model.

The same five splits between train and test respondents are used as in the hyperparameter optimization. Therefore, we can benchmark the CNN with activity cycling with the results in Table 4.3. Compared to the accuracy of the CNN, both LSTM models performed less good. Although RF contains fewer parameters, it is the best performing model. This might be due to the different features. Comparing feature sets could be a topic for further research.

[3] Scikit-learn version 0.24.2, Tensorflow version 2.4.1.

Table 4.3 Accuracy of four models, together with balanced accuracy and the corresponding standard deviations. Accuracy is abbreviated as acc. and standard deviation as std.

	Acc.	Std. acc.	Balanced acc.	Std. balanced acc.
CNN	0.924	0.044	0.866	0.041
1-layer LSTM	0.869	0.058	0.736	0.062
2-layer LSTM	0.856	0.057	0.764	0.058
RF classifier	0.970	0.011	0.940	0.021



True \ Prediction	cycling	walking	running	jumping	standing	climbing stairs	sitting
cycling	6097	5	3	0	0	1	1
walking	0	3093	0	0	0	15	0
running	0	7	2975	0	0	2	0
jumping	0	0	0	300	41	1	0
standing	0	0	0	5	2760	6	271
climbing stairs	0	14	3	0	6	425	0
sitting	0	0	1	0	10	0	3077

Figure 4.4 Confusion matrix of Random Forest

Figure 4.4 shows the confusion matrix. This shows that the performance of the Random Forest is comparable to the performance of the Neural Network. Table 4.4 shows that only the recall for activity jumping is less than 0.9. Recall and precision can be calculated from the confusion matrix. The performance of the Random Forest on the week data drops dramatically, as with the neural network, with an accuracy of 0.41 and a balanced accuracy of 0.25. There is a big difference in accuracy between participants, ranging from 0.17 to 0.80.

Table 4.4 Classification report of Random Forest.

	precision	recall	F1-score	support
cycling	1.00	1.00	1.00	6107
walking	0.99	1.00	0.99	3108
running	1.00	1.00	1.00	2984
jumping	0.98	0.88	0.93	342
standing	0.98	0.91	0.94	3042
climbing stairs	0.94	0.95	0.95	448
sitting	0.92	1.00	0.96	3088
accuracy			0.98	19119
macro avg	0.97	0.96	0.97	19119
weighted avg	0.98	0.98	0.98	19119

4.5 Weekdata

The previous paragraphs described the performance of the activity recognition models during the laboratory sessions. In the end, Statistics Netherlands would like to create a model that is able to extrapolate well to the free living data. Therefore, we aimed to predict periods in the weekdata with a model trained on the labdata. We first evaluate the performance of a model trained on labdata. The weekdata contains more activities than the labdata. The activities in the weekdata are also different in duration and more diverse. For example, respondents can stand in the kitchen (alternated with walking) or work while standing. Therefore, a model based on labdata is not likely to generalize well to the weekdata and we will try to augment the training data by using selections of the weekdata.

Although the respondents filled in a diary, we do not know all details about the activities during the free living week. The evaluation criteria (accuracy etc.) are, therefore, evaluated with respect to activPAL labels. The activPAL classification label sedentary can be compared with the predictions of sitting, and activPAL labels non-wear and travelling are not present in the labsessions. These labels were removed before calculating the metrics. The diary activities sleeping, sports, travelling and non-wear were removed with a margin of 15 minutes before and after the activity, since the model based on the laboratory activities would not be able to classify these activities well. Non-wear in the diary can refer to taking off one or both attached sensors, but since there is enough weekdata, we decided to remove all non-wear periods. Diary activities cycling and working are not removed, since the model is trained to recognize cycling and working does likely contain activities like sitting, standing and walking.

Figure 4.5 shows a comparison between activPAL's labels and the most frequently predicted class for every minute. It turns out that the model trained on labdata often incorrectly predicts cycling, resulting in an accuracy of 0.114 and a balanced accuracy of 0.154. Further research is necessary to analyse this behavior. It is remarkable that the model predicts only a few minutes of running. Recall that sport periods as noted in the diary are removed, and it is hardly likely for running to occur during other periods. The performance differs a lot between the respondents, with a maximal overall accuracy of 0.68.

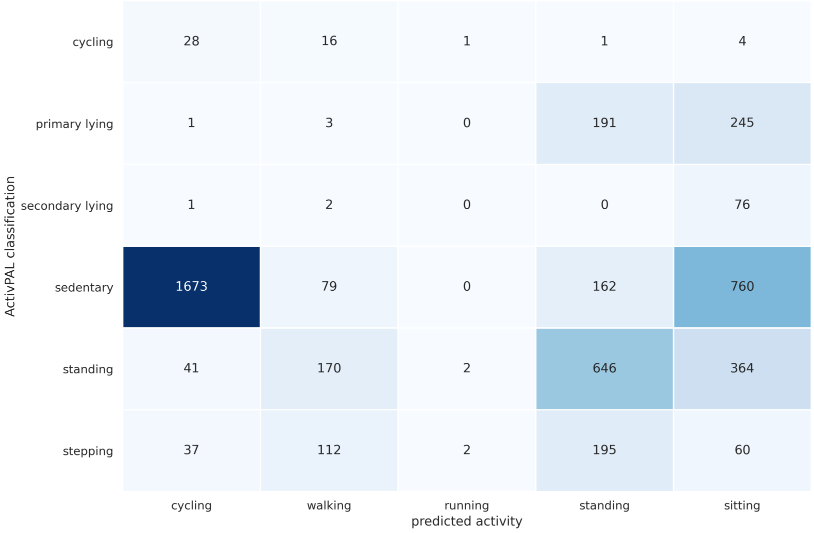


Figure 4.5 Confusion matrix of ActivPAL predictions and the predictions of the model trained on labdata during the free living week.

4.6 Short summary of section

In this section, we described the performance of machine learning models trained on measurements of different sensors to answer the second research question: How well can machine learning models classify activities based on labdata? The neural network trained on ActivPAL's accelerometer data can quite accurately classify the time series, followed by IMU tibia. However, the model confuses cycling light with cycling heavy. Adding heart rate as an additional feature increases the performance of the models. Hyperparameter optimization did not have a large impact on the performance. A neural network is not the only type of model that can accurately classify the labdata, a random forest model performed slightly better. The train-test split of respondents does affect the performance. This behaviour might be caused by the differences between the respondents (e.g. general activity level, sex) or between the execution of the activities (e.g. did respondents follow the exact instructions?). Recall that we trained the models on a relatively small sample size of 36 respondents.

The third research question is: How well can machine learning models classify activities based on weekdata? First, a model trained on labdata is used to classify the weekdata. The results are compared with ActivPAL's classification since the activities in the diary are different from the laboratory activities. Even though ActivPAL's classification contains different classes, we can conclude that the quality of the classification is very poor. Cycling is for example often classified as sedentary. Experiments to supplement the training data with measurements from specific periods from the free living period do not seem very promising at this moment. Semi-supervised machine learning might be more promising when the quality of the classification is better.

5 User Friendliness

An important aspect of the choice for an accelerometer for large scale use is user friendliness. In contrast to dedicated research projects, the general public will be less inclined to humour the researcher and will be apt to reject the device or stop prematurely if the device is complicated to mount or use. Investigating user friendliness was no explicit goal of the small scale research project we report upon here. Nevertheless, some observations can be made. Specifically, the number of non-wear hours of the two free living devices can be monitored, and the remarks participants made spontaneously.

The activPAL was worn for a means of 7 days (21 of the participants wore the device for 8 days, 28 wore them the entire 7 days); four participants did not wear it the full 7 days for reasons that had nothing to do with usability (we needed the devices back for another purpose). One participant complained about skin irritation but continued wearing the device. Adherence to the UKK protocol was slightly less. The UKK had to be taken off during showering or other activities where the device risked getting wet. Also, the device had to be transferred from waist to pulse during sleeping. Taking the device off bears the risk that participants forget to put it back on.The UKK was worn 7 or 8 days by 27 participants. 5 participants did not wear the UKK every day. One participant wore the UKK only 3 days. For one participant there were no data. 3 participants wore the device for less than 12 hours for at least one day.

In a subsequent pilot among 47 participants where the activPAL was used (Toepoel et al., 2021), user evaluation was an explicit part of the study. Participants were asked how they evaluated wearing the activPAL in an open question. Regretfully, only 26 participants answered the question. Only one participant was negative, finding the adhesive tape unpleasant and itchy, and the manual to mount the device unclear. 18 participants were unequivocally positive ('you never even notice that you are wearing it'), while 7 participants were positive, but with a qualification that some aspects were less than optimal (removing the adhesive plaster from a hairy leg, and also itchiness and skin irritation). Part of the skin irritation was not caused by the adhesive plaster, but by the latex finger condom that was used to waterproof the activPAL. This can be remedied by using other materials. These findings on respondent evaluation corroborate other research, (e.g., Berendsen et al., 2014; Edwardson et al., 2017). In this study, the mean number of valid days was 6,5, with 78 percent of participants wearing the device for the requested 7 days.

The sensors that were only used in the lab were not evaluated explicitly on user friendliness. It was evident from the start that most would not be feasible for use outside the lab situation, either because the machine was too large (the Vyntus) or because the device was too expensive (the Hexoskin). The latter was also impractical, as participants would either have had to wear the Hexoskin vest day and night for seven days, or had to have at least one spare one. The fact that the device needs in practice to be mailed out and also returned by mail is an additional argument to strive for a small device. The IMUs worn on the shin and arm were neither very practical, as they were both quite large, and also came without holding device; they had to be taped on the leg and arm. As a result, the IMUs were prone to get loose easily.

6 Summary and discussion

In this paper we describe a pilot aimed at selecting a sensor system for use in subsequent physical activity research. The ultimate aim is to use objective sensor data in large scale health surveys, to replace or supplement survey questions on physical activity.

In a lab setting 40 participants in four categories (male or female and active or not active) performed a number of tasks (sitting, standing, stepping at two intensity levels, bicycling at two intensity levels, walking stairs and jumping), observed by trained lab assistants who noted the precise start and end times of each task. Participants' metabolic response was measured through breath-by-breath analysis by Vyntus™ CPX, allowing the precise calculation of energy expenditure. These measurements served as the benchmark for accelerometer data from four different sensor systems, mounted on various parts of the body: the thigh worn activPAL, the UKK worn on the waist (and pulse during sleeping), a IMU on the shin and one on the pulse, and a Hexoskin shirt. The Hexoskin additionally measured heart rate. In a subsequent week, participants wore the AP and the UKK in a free living setting (i.e., at home and at work), while keeping a diary of their physical activity. In the diary the start and end times of the activities 'sleeping', 'working', 'travelling', 'bicycling', 'doing sports', and 'non-wear' were noted.

6.1 Intensity

The first research question was: How can we estimate intensity based on the lab- and weekdata? We compared three metrics, heart rate, MET and Median Amplitude Deviation (MAD), using the measurements of the laboratory sessions. All three metrics have disadvantages. MAD is the best out of the three metrics to measure intensity. Kuppevelt et al., 2019 investigate yet another approach to classify the measurements into intensity categories, unsupervised machine learning, that seems promising. In this approach, activities are constructed based on free living observed data directly, without the need for calibration studies. Activities are inferred from the distributions of observations and their duration. The search for the optimal method is still ongoing, however.

Heart rate by itself is not sufficient to determine the intensity of activities; although there are participants where the correlation between VO_2 and heart rate is almost perfect, there are also participants where the correlation is very low (.27). Some participants' heart rates were lower during cycling than during standing for example, making it difficult to estimate intensity. Heart rate could give valuable additional information when added to an accelerometer. However, adding a second device might be an extra burden for participants, and also an additional strain on funds. A heart rate measurement would be valuable in future testing to develop free living algorithms, but for structural fieldwork for population surveillance, it should be avoided if it would mean adding a second device.

The MET measure, the Metabolic Equivalent of Task, is approximately 3.5mL VO_2/kg/minute and represents the intensity of quietly sitting. MET values of the lab activities could be determined by the Vyntus measurements. The SQUASH questionnaire and the activPAL determine energy expenditure by assigning MET values to activities. The SQUASH inventories the activities like walking, biking, gardening, cleaning, but also the precise nature of sports. The activPAL distinguishes a more limited activity palette; only sitting, standing, stepping, and biking. ActivPAL first determines the activity, and subsequently assigns the appropriate MET value.

The third way we looked into intensity was by studying the Median Amplitude Deviation. This is the measure of intensity used by UKK and is an established way to determine intensity (see e.g., Kuppevelt et al., 2019). Low, moderate and vigorous intensity are subsequently determined by choosing cut-off points. Those cut-off points are different for different populations (children, adults, elderly). The cut-off points in this pilot were determined by the lab results of one participant. The MAD is substantially better in distinguishing activities than heart rate. It would be interesting to compare the MAD and heart rate after a correction for the baseline heart rate. The MAD values correlated moderately highly with VO_2 for the lab activities. For the lab and week data, low and moderate MAD intensity coincides very well with the activPAL classifications: the low intensity activities of lying and sitting, as well as non wear were classified as such for 92 to 100% of cases. Stepping and bicycling were classified as moderate of intensive acitivity in 97 and 99% of cases, respectively. There were two classes with low agreement: standing and travelling, that were classified as moderately intensive by MAD in 70% and 61% of cases, respectively. It is unclear at this point where this confusion stems from: is activPAL incorrectly determining standing, are the cut-off points too liberal, was the one participants on whose values the cut-off points were determined perhaps not representative for all the participants?

Comparison with participants' diaries has determined that activPAL not in all cases correctly determined travelling. Car and train travel were mostly correctly recognized as travel, but travelling by bus and tram were not. The MAD calculation seems to suffer from the same over-estimation of activity.

6.2 Activity classification of labdata

The second research question was: How well can machine learning models classify activities based on labdata? For each sensor a model was trained to classify activities. All models worked relatively well. ActivPAL, worn at the upper thigh, achieved the best accuracy. The IMU, worn on the tibia, performed the second best even though the training set was smaller. All models benefit from adding the heart rate as an additional feature. However, the added specificity of heart rate was not very large: the accuracy of the classification models based on activPAL data increased with 0.034 when heart rate was added. It would be interesting to include derived features next to heart rate such as frequencies that are used by the Random Forest and analyse the most important features of the model in more detail. It must be stressed that the UKK results were not included in the comparison: UKK data were not available in the raw format needed to link data to the Vyntus and to the Hexoskin's heart rate measure. Analysis of the separate UKK showed that UKK was not sufficiently able to recognise bicycling (Bikker et al., 2020), a finding that is recognized by other research on hip worn accelerometers (Tarp et al., 2015). In addition, the fact that no raw data could be received from the UKK, makes the UKK not suitable for use for CBS.

Note that even although our models use only accelerometer data, the frequency and domain of the measurements differed over the sensors. Therefore, the choice for activPAL does not necessarily mean that the upper thigh is the best location to wear a sensor for activity recognition in general, although literature supports the location of the upper thigh as the favoured position (e.g., Kozey-Keadle et al., 2012, Kuster et al., 2020 for monitoring sedentary behaviour; see also Stamatakis et al., 2019). Additionally, information on inclination that can be calculated from the angles between the vectors, as yet not included in our models, will potentially increase accuracy of activity recognition.

6.3 Activity classification of weekdata

The third research question was: How well can machine learning models classify activities based on week data? While the machine learning algorithms were able to reliably classify the activities performed in the lab, the algorithms very poorly generalised to the free living period. Despite the fact that most activities people do during the day exist of lying, sitting, standing, and stepping, the model's accuracy was 0.114. Especially cycling was often wrongly classified. This might be due to incorrect and inaccurate diary keeping, both in classification as in timing the performed activity. In three subsequent pilots that were performed since, participants were not asked to keep a diary, as a result of the very limited use they were in this pilot. However, accurately determining the degree of physical activity during the free living period is the ultimate goal. Adherence to physical activity norms and insight into people's sedentary behaviour is calculated on the basis of the free living week. There are multiple arguments why the model does not extrapolate well. We will list the most important arguments here.

- The movements in the laboratory are way simpler and less volatile than real life movements. For example, hikers have to stop for traffic lights in real life and bicyclers encounter speed bumps that will influence the acceleration. House cleaning is for example composed of short movements like standing, walking, picking up stuff.
- Some activities in real life do not categorise well into the eight activity classes we measured in the lab. Relatively static sports like boxing or weight lifting will not be classified as intensive activity when measured with an accelerometer. On the other hand, while travelling, movement of the vehicle can be incorrectly classified as movement of the participant. This was noted when comparing the activPAL classification with diary information.
- The length and type of activities varies in real life over the respondents. Some activities might not even occur for some respondents. Some activities occur for example more often during the weekend or working hours. During the laboratory sessions, all respondents performed the exercises for a fixed time.

6.4 User friendliness

The ultimate goal of this research project is to study the feasibility of using accelerometers for objectively measuring physical activity in population surveillance. A prerequisite to this end is not only that we succeed in deriving the desired statistics, but also that we are able to convince the public to wear the device, in the manner that we wish, and to continue wearing it. The preliminary findings show that most participants are able to mount the activPAL according to specification, that adherence to protocol is high and that the device is not seen as a burden to wear. It has to be specified, however, that our experiences so far are limited to dedicated participants who either volunteered, or were members of a panel. Future studies among the general population are foreseen where we can more realistically gauge, and possibly influence, people's willingness and adherence.

6.5 Lessons Learned

An important aspect of this pilot was gaining experience with the data science aspects of accelerometer measurement. Important lessons were learned on the topic of data cleaning of these kinds of data, and the various machine learning approaches that could be used.

Measurements of 36 of the 40 respondents in the small scale project were successfully cleaned and aligned with the start- and end time of registered activities. Synchronising the timestamps of the sensor with the timestamps of the activities turned out to be challenging. We recommend to precisely monitor the start and end time of activities. Calibration of sensors before or after use needs to be part of the protocol. This will likely reduce the number of hours to clean the data after getting the measurements from the sensors. Matching the data of the different sensors was also time consuming, since frequency and the range of the measurements was different. Vyntus saved for example only data during the activities, and used an index number instead of a timestamp.

Several machine learning techniques were applied and tested: Convolutional Neural Networks, Long-Short Term Memory models with either one or two LSM layers, and simple Random Forest classifiers. All the models were tweaked by fine tuning the hyper parameters and by trying different splits between training and test data. Although the CNN and LSTM are specifically suitable for time series data, the simple RF model achieved a slightly higher accuracy in classifying activities. A master student looked into the possibility of using Multilevel Hidden Markov models for the free living data (Vogelzang, 2021). Still other options could be considered, like training a hierarchical model for activity recognition. However, we felt that it would not be worthwhile to pursue perfecting the models on imperfect and insufficient data. Since the models classify short periods of measurements, the performance of the models is highly dependent on the timestamps of the activities. We stress that the accuracies as found in this study can, therefore, easily be improved by follow up studies.

6.6 Conclusion

Given the performance of the models, the costs, flexibility of analysis, and response and usability aspects like adherence to protocol, the activPAL is our sensor of choice. Adding a heart rate monitor would increase the accuracy of measuring intensity. Also measurement of GPS location could be very informative of the speed with which, and the locality where (home, office, sports accomodation), people move. However, a careful consideration needs to be made if this increase in accuracy outweighs the increased costs and respondent burden of a second sensor. Additional research is necessary to balance these requirements.

We have indicated that at least one additional pilot is necessary with the chosen solution for us to understand physical activity in the free living period. Ideally, the pilot would have more respondents than the present one, to be able to incorporate differences in for example age, gender, general activity level in activity recognition models. In general, however, more detailed research with additional sensors on different body parts and positions would be a valuable addition to the knowledge base.

To increase the accuracy of the analysis of free living data, we propose three tracts: the first is an additional qualitative test, where participants are asked to fill in a complete 24 hour diary of all their activities, in 10 minute bouts. Linking to the Time Use survey app under development would be ideal. The goal of such a test would be to build a training pool of more varied and fully labelled activities. In these analyses, inclination will be used as a source of information, in addition to acceleration. Training the models with transitions between activities is another promising direction for future research, especially in combination with allowing respondent heterogeneity in model parameters. Adding more information about the context of the movement, like time of the day, week or weekend, working hours, or the weather might also help to increase the performance. However, it will continue to be difficult to recognise various sports with various intensities, something that needs to be inventoried on behalf of the Physical Activity Guidelines.

The second tract is to move away from attempts to classify all behaviours, and instead focus more on classifying the intensity of behaviours, in a sense the more traditional approach. The activity guidelines specify that people should be moderately active during a number of minutes (or vigorously active during a shorter time). Recognising intensity of activity is perhaps easier than recognising each and every activity as such. This is a tract that is also chosen by e.g., Kuppevelt et al, (2019). In this paper we have studied ways to determine intensity, by looking into heart rate, MET values, and median amplitude deviation (MAD) of the three accelerometer axes. Focusing on intensity, however, precludes the possibility to one day be able to distinguish between a broad set of different types of sports in accelerometry, the other dimension of the activity guidelines.

The third tract is to combine the previous tracts in a hybrid estimation method. Respondents can fill a diary with start and end times of their activities, possibly complemented with other relevant information. The diary should be designed such that the respondent can easily list their activities. It is also possible to experiment with more general classes such as work, sports, travelling, getting out of bed. Next to the diary, sensor measurements can be used to estimate the intensity of activities. The intensity and activity class can then be combined to get a more accurate estimate of the movements of the respondent. One could use diary information to estimate a respondent and time dependent prior distribution. This prior can for example be used to map the output of the neural networks to a classification per minute or as the transition prior of an Hidden Markov Model as used by Vogelzang, 2021. Again, the benefits to quality and accuracy of adding a diary in fieldwork need to be carefully balanced with respondent burden.

We need to be able to analyse these data independently, without having to resort to the proprietary software developed by activPAL. For the time being, the activPAL algorithms may function as a benchmark to gauge the performance of our own algorithms. ActivPAL is used by a large number of researchers of physical activity, who are active in research communities like ProPASS, and who develop and share algorithms. Also, by using the activPAL we do not need to do the extensive research necessary to prove that the activPAL is able to accurately classify activity, others have done that for us. Additionally, using devices that are used by several research institutes, also in the Netherlands, makes it possible to share resources and know how, as is presently done with AmsterdamUMC and Maastricht University.

We set out this research in search of more objective and more precise measurement of physical activity than the SQUASH questionnaire that is used presently. It is known from literature that (respondents in) the SQUASH tend to overestimate their physical activity, a finding that is replicated in one of our own pilots where we compare the SQUASH with the activPAL and people’s own accelerometer (Toepoel et al., 2021). We have found in this first pilot that accelerometers also have their faults, and may for example classify riding a bus as moderately intensive activity. In addition, the choice of epoch length and the determination of the cut-off point can all impact the measured intensity. Careful handling of these parameters is necessary in order to measure reliably (Brondeel, 2016).

The experiments that are presently being performed, the additional qualitative pilot that is suggested above, and a larger scale field test that is envisaged in the near future will shed more light on how to balance these respective faults. It is very well probable that some form of questionnaire in addition to sensor measurement will be necessary to reliably estimate physical activity in all its forms.

References

Ainsworth, B. E., Haskell, W. L., Whitt, M. C., Irwin, M. L., Swartz, A. M., Strath, S. J., O Brien, W. L., Bassett, D. R., Schmitz, K. H., Emplaincourt, P. O., et al. (2000). Compendium of physical activities: An update of activity codes and met intensities. Medicine and science in sports and exercise, 32(9; SUPP/1), S498–S504.

Berendsen, B. A., Hendriks, M. R., Meijer, K., Plasqui, G., Schaper, N. C., & Savelberg, H. H. (2014). Which activity monitor to use? validity, reproducibility and user friendliness of three activity monitors. BMC Public Health, 14(1), 749.

Bikker, R., Heuvel van, G., Hoek van, S., Hoogland, J., Lodder, B., Windmeijer, D., Luiten, A., & Schouten, B. (2020). The accelerometer project [unpublished].

Brage, S., Brage, N., Franks, P. W., Ekelund, U., Wong, M.-Y., Andersen, L. B., Froberg, K., & Wareham, N. J. (2004). Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure. Journal of applied physiology.

Brondeel, R. (2016). The relevance of transport to promote physical activity : Addressing challenges related to the measurements and the observational analysis of transport-related physical activity, and the simulation of shifts in transportation mode (Doctoral dissertation). Université Pierre et Marie Curie - Paris VI.

Edwardson, C. L., Winkler, E. A. H., Bodicoat, D. H., Yates, T., Davies, M. J., Dunstan, D. W., & Healy, G. N. (2017). Considerations when using the activpal monitor in field-based research with adult populations. Journal of Sport and Health Science, 6(2), 162–178.

Esliger, D. W., & Tremblay, M. S. (2007). Establishing a profile of physical activity and inactivity: The next generation. Applied physiology, nutrition, and metabolism= Physiologie appliquee, nutrition et metabolisme, 32, S217–30.

Ferrari, P., Friedenreich, C., & Matthews, C. E. (2007). The role of measurement error in estimating levels of physical activity. American journal of epidemiology, 166(7), 832–840.

Freedson, P., Melanson, E., & Sirard, J. (1998). Calibration of the computer science and applications, inc. accelerometer. Med Sci Sport Exer, 30(5), 777–781.

Fruin, M. L., & Rankin, J. W. (2004). Validity of a multi-sensor armband in estimating rest and exercise energy expenditure. Medicine and science in sports and exercise, 36(6), 1063–1069.

Helmerhorst, H. H. J., Brage, S., Warren, J., Besson, H., & Ekelund, U. (2012). A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. International Journal of Behavioral Nutrition and Physical Activity, 9(1), 1–55.

Kozey-Keadle, S., Libertine, A., Staudenmayer, J., & Freedson, P. (2012). The feasibility of reducing and measuring sedentary time among overweight, non-exercising office workers. Journal of Obesity, 2012, Article ID 282303.

Kraakman, R. (2021). The validity of consumer-level activity trackers in research on physical activity: A comparison with self-report and a professional accelerometer on representation and measurement (Master's thesis). Utrecht University.

Kuppevelt, D. v., Heywood, J., Hamer, M., Sabie, S., Fitzsimons, E., & van Hees, V. (2019). Segmenting accelerometer data from daily life with unsupervised machine learning. PLOS one, 14(1).

Kuster, R. P., Baumgartner, D., Hagströmer, M., & Grooten, W. J. A. (2020). Where to place each sensor to measure sedentary behaviour? a method development and comparison among various sensor placements and signal types. Journal for the Measurement of Physical Behaviour, 3, 274–284.

Londeree, B. R., & Moeschberger, M. L. (1982). Effect of age and other factors on maximal heart rate. Research quarterly for exercise and sport, 53(4), 297–304.

Luiten, A., & Voermans, R. (2019a). Protocol assistent 1 for one participant at a time (in dutch: Stappenplan assistent 1 bij één respondent tegelijk) [unpublished].

Luiten, A., & Voermans, R. (2019b). Protocol assistent 2 for one participant at a time (in dutch: Stappenplan assistent 2 bij één respondent tegelijk) [unpublished].

Matthews, C. (2005). Calibration for accelerometer output for adults. Med Sci Sport Exer, 37(11), S512–S522.

Nicolaou, M., Gademan, M., Snijder, M., Engelbert, R., Dijkshoorn, H., Terwee, C., & Stronks, K. (2016). Validation of the squash physical activity questionnaire in a multi-ethnic population: The helius study. PLoS One, 11(8), e0161066.

Prince, S. A., Adamo, K. B., Hamel, M. E., Hardt, J., Gorber, S. C., & Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: A systematic review. International journal of behavioral nutrition and physical activity, 5(1), 1–24.

Sallis, J. F., & Saelens, B. E. (2000). Assessment of physical activity by self-report: Status, limitations, and future directions. Research quarterly for exercise and sport, 71(sup2), 1–14.

Sasaki, J., John, D., & Freedson, P. (2011). Validation and comparison of actigraph activity monitors. J Sci Med Sport, 14(5), 411–416.

Schouten, B., & Voermans, R. (2019). Projectplan beweegmeter [unpublished].

Shephard, R. J. (2003). Limits to the measurement of habitual physical activity by questionnaires. British journal of sports medicine, 37(3), 197–206.

Stamatakis, E., Koster, A., Hamer, ., Mark, & Holtermann, A. (2019). Emerging collaborative research platforms for the next generation of physical activity, sleep and exercise medicine guidelines: The prospective physical activity, sitting, and sleep consortium (propass). British Journal of Sports Medicine, 54.

Staudenmayer, J., He, S., Hickey, A., Sasaki, J., & Freedson, P. (2015). Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. Journal of Applied Physiology, 119(4), 396–403.

Tarp, J., Andersen, L., & Østergaard, L. (2015). Quantification of underestimation of physical activity during cycling to school when using accelerometry. Journal of Physical Activity and Health, 12, 701–707.

Toepoel, V., Luiten, A., & Zandvliet, R. (2021). Response, willingness, and data donation in a study on accelerometer possession in the general population. Survey Practice, 14(1).

Troiano, R., Berrigan, D., Dodd, K., Mâsse, L., Tilert, T., & Mcdowell, M. (2008). Physical activity in the united states measured by accelerometer. Med Sci Sport Exer, 40(1), 181–188.

Troiano, R., & Freedson, P. S. (2010). Promises and pitfalls of emerging measures of physical activity and the environment. American journal of preventive medicine, 38(6), 682.

Vähä-Ypyä, H., Husu, P., Suni, J., Vasankari, T., & Sievänen, H. (2018). Reliable recognition of lying, sitting, and standing with a hip-worn accelerometer. Scandinavian journal of medicine & science in sports, 28(3), 1092–1102.

Vähä-Ypyä, H., Vasankari, T., Husu, P., Mänttäri, A., Vuorimaa, T., Suni, J., & Sievänen, H. (2015). Validation of cut-points for evaluating the intensity of physical activity with accelerometry-based mean amplitude deviation (mad). PloS one, 10(8), e0134813.

Vähä-Ypyä, H., Vasankari, T., Husu, P., Suni, J., & Sievänen, H. (2015). A universal, accurate intensity-based classification of different physical activities using raw data of accelerometer. Clinical physiology and functional imaging, 35(1), 64–70.

van Hees, V., Fang, Z., Langford, J., Assah, F., Mohammad, A., da Silva, I., Trenell, M., White, T., Wareham, N., & Brage, S. (2014). Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: An evaluation on four continents. J Appl Physiol, 117, 738–744.

Vogelzang, J. (2021). Multilevel hidden markov model for activity recognition (Master's thesis). Utrecht University.

Ward, D., Evenson, K., Vaughn, A., Rodgers, A., & Troiano, R. (2005). Accelerometer use in physical activity: Best practices and research recommendations. Medicine and science in sports and exercise, 37, S582–8. https://doi.org/10.1249/01.mss.0000185292.71933.91

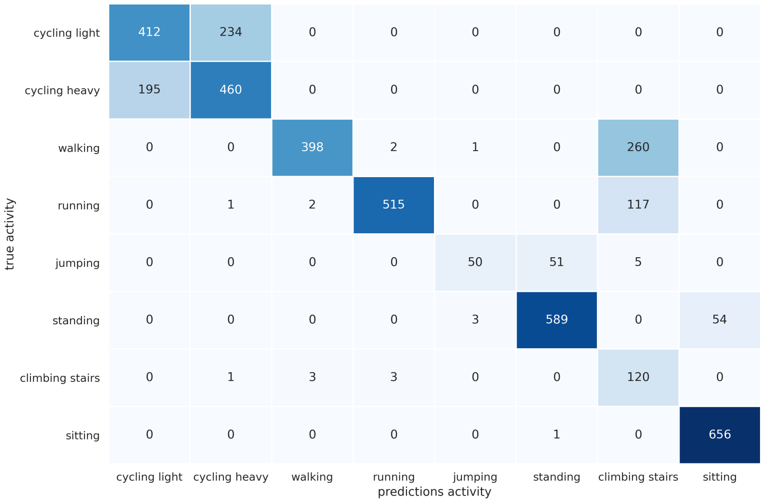
Welk, G. J., Kim, Y., Stanfill, B., Osthus, D. A., Calabro, A. M., Nusser, S. M., & Carriquiry, A. (2014). Validity of 24-h physical activity recall: Physical activity measurement survey. Medicine and science in sports and exercise, 46(10).

Wendel-Vos, G. W., Schuit, A. J., Saris, W. H., & Kromhout, D. (2003). Reproducibility and relative validity of the short questionnaire to assess health-enhancing physical activity. Journal of clinical epidemiology, 56(12), 1163–1169.

Wijndaele, K., Westgate, K., Stephens, S. K., Blair, S. N., Bull, F. C., Chastin, S. F., Dunstan, D. W., Ekelund, U., Esliger, D. W., Freedson, P. S., et al. (2015). Utilization and harmonization of adult accelerometry data: Review and expert consensus. Medicine and science in sports and exercise, 47(10), 2129.

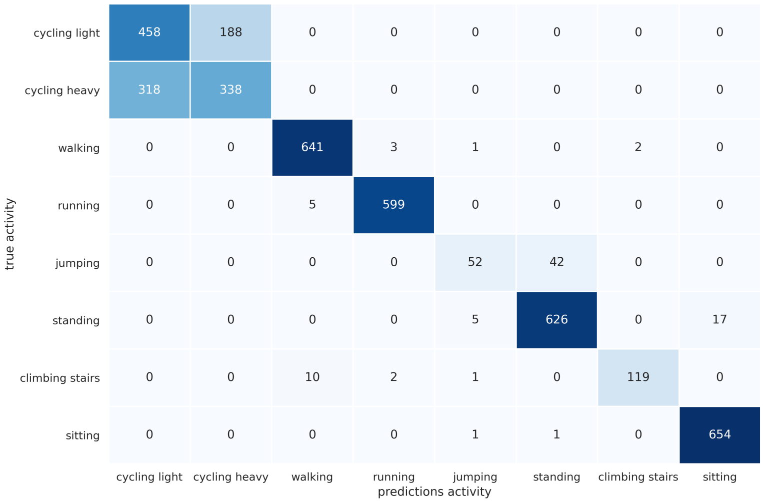
Appendix

A Activity recognition



	cycling light	cycling heavy	walking	running	jumping	standing	climbing stairs	sitting
cycling light	412	234	0	0	0	0	0	0
cycling heavy	195	460	0	0	0	0	0	0
walking	0	0	398	2	1	0	260	0
running	0	1	2	515	0	0	117	0
jumping	0	0	0	0	50	51	5	0
standing	0	0	0	0	3	589	0	54
climbing stairs	0	1	3	3	0	0	120	0
sitting	0	0	0	0	0	1	0	656

Figure A.1 Confusion matrix of the model trained on the first train-test split with optimized hyperparameters. The values in the matrix correspond to the number of observations in the test data (without overlapping slices).



	cycling light	cycling heavy	walking	running	jumping	standing	climbing stairs	sitting
cycling light	458	188	0	0	0	0	0	0
cycling heavy	318	338	0	0	0	0	0	0
walking	0	0	641	3	1	0	2	0
running	0	0	5	599	0	0	0	0
jumping	0	0	0	0	52	42	0	0
standing	0	0	0	0	5	626	0	17
climbing stairs	0	0	10	2	1	0	119	0
sitting	0	0	0	0	1	1	0	654

Figure A.2 Confusion matrix of the model trained on the second train-test split with optimized hyperparameters.

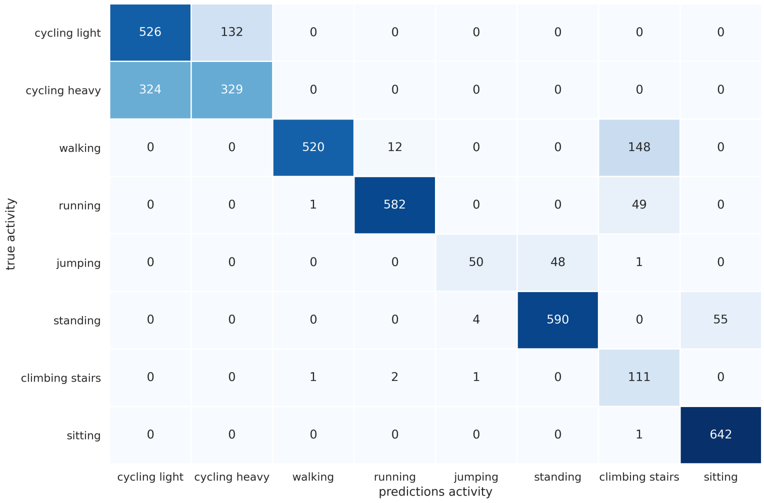


Figure A.3 Confusion matrix of the model trained on the third train-test split with optimized hyperparameters.

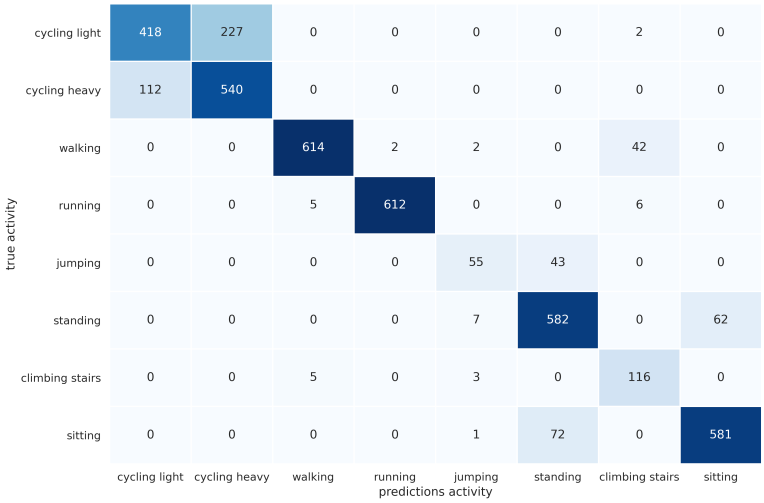


Figure A.4 Confusion matrix of the model trained on the fourth train-test split with optimized hyperparameters.

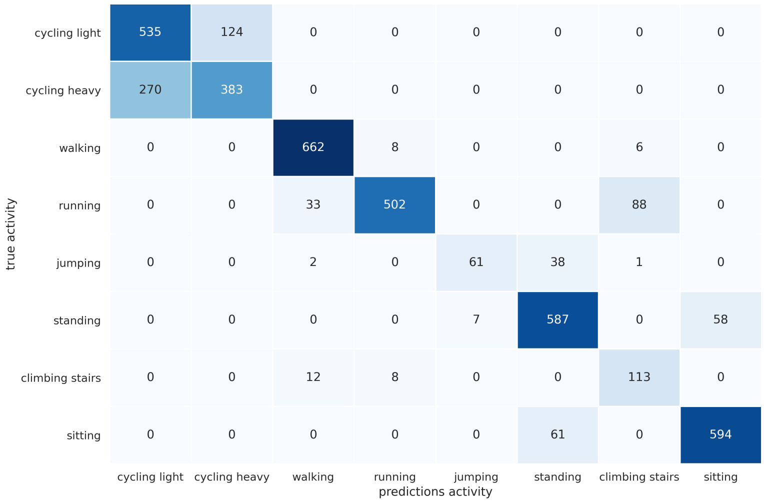


Figure A.5 Confusion matrix of the model trained on the fifth train-test split with optimized hyperparameters.

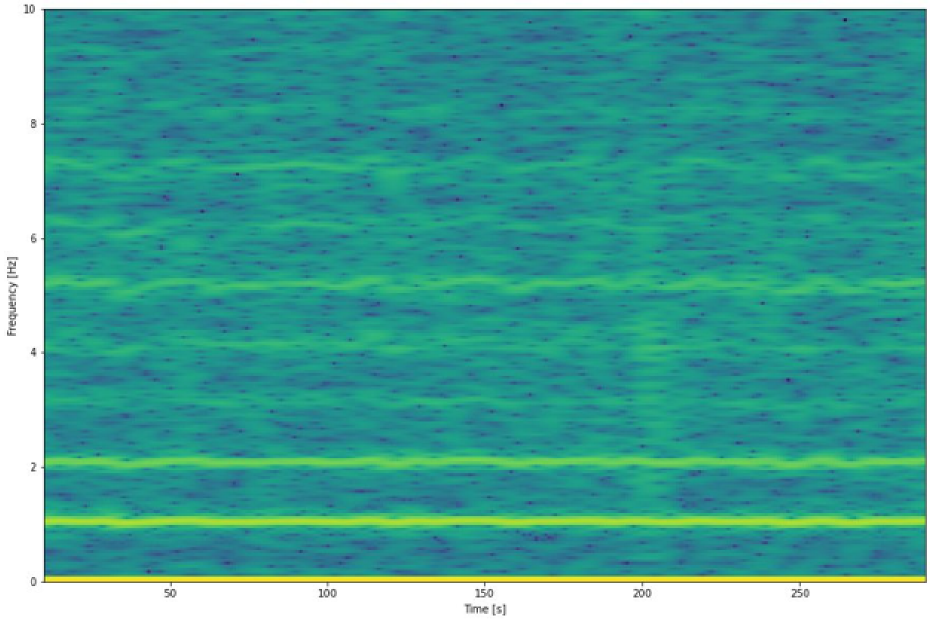


Figure A.6 Spectogram of one respondent during a laboratory session. Top frequencies are shown in yellow, in contrast to less often used frequencies that are shown in blue.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information