Universiteit Leiden
The Netherlands

Bachelor Computer Science

Energy expenditure estimation for wheelchair users using activity type classification

Marc Boel

Supervisors:
Iris Yocarini & Stylianos Paraschiakos

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

22/07/2021

Abstract

Among wheelchair users, cardiovascular disease is very common. This is mainly because of their lower daily energy expenditure (EE), which is a result of less physical exercise. A good EE estimation method could help wheelchair users with their exercise and researchers with health studies. Current research on EE estimation for wheelchair users is still limited but active, and progress in EE estimation that worked for non-wheelchair users is actively being converted to work for wheelchair users. In this thesis, an activity-specific EE estimation method was developed based on random forests, which first predicts the activity type and uses that information to help estimate EE for spinal cord injury (SCI) and lower limb amputation (LLA) patients. Two different activity type models were assessed, a fine grained approach using twelve activities and a broader approach using three activity levels. To assess a baseline score, a basic random forest regression model was used, which had a MAPE of 0.367 in the SCI group. Although multiple versions showed insignificant performance differences, the activity-specific model that used three classes and was trained on the actual activities showed improvement with a MAPE of 0.350. The LLA group was too small to show any significant performance differences. Overall, we showed that an activity-specific EE estimation method can bring improvements, although only minimal, and further research is needed before any practical use.

Contents

1 Introduction

For wheelchair users, cardiovascular disease is a major problem. For example, cardiovascular disease is the leading cause of mortality for people with spinal cord injuries, compared with the able-bodied population [GKC+05]. When someone does less physical exercise, and thus has a lower energy expenditure (EE), the chances of contracting cardiovascular disease become much higher. According to Myers et al. [MLK07] the risks include a greater change of contracting conditions like obesity, lipid disorders, metabolic syndrome, and diabetes. This is mainly because the daily EE is significantly lower in patients with spinal cord injury (SCI) and since it is difficult for many wheelchair users to get the needed amount of proper exercise, this is a big issue. The reason for the lower amount of exercise in wheelchair users is not only because of a lack of motor function, but also because of a lack of accessibility and fewer opportunities to engage in physical activity. It has been shown that physical activities (PA) can reduce these risks and help improve the quality of life for these patients [SMGT13], so it is important to promote PA specifically for this population.

For this purpose, it is important to improve accessibility and opportunities for physical activities for wheelchair users by improving PA monitoring techniques. Monitoring PA's can help patients with their exercise by giving them more insight into their daily activities and exercise, but it can also, among other things, help researchers or trainers decide which training schedules work and which do not. Despite all of this, PA monitoring support for wheelchair-bound individuals is mediocre at best and many wheelchair users choose not to use them [CCMH15]. The aim of this study was to improve EE estimation based on accelerometer data to improve PA monitoring for wheelchair users so it can become more useful.

Commercially available wearables help able-bodied individuals with physical exercise by providing EE estimation methods. These methods make use of supervised regression algorithms that can use the sensors of the wearable to estimate the EE of the user. EE can be roughly divided into two categories: *Resting* EE (sometimes called passive EE) and *active* EE. Resting EE is the amount of energy that the body uses to stay alive. This includes the digestive system, circulatory system and respiratory system, but also our brain, cell growth and controlling body temperature. This is also referred to as the *basal metabolic rate*. Active EE is the energy that is used to voluntarily move the body. This includes both exercise and non-exercise. The daily EE is the sum of the total resting EE and active EE of a whole day. This thesis focuses on both resting and active EE.

Wearables usually contain heart rate sensors and accelerometers to detect precise movement, but in some cases also more advanced sensors e.g. blood pressure sensors or blood oxygen level sensors. This data is then processed into information that is easier to interpret by humans, like EE or the activity type. More advanced wearables can even help the user by recommending when to exercise, rest, eat or hydrate, which are estimations made by algorithms based on the data that is captured by the sensors. These wearables unfortunately do not work well for people in wheelchairs, because they often over- (or in some case under-)estimate EE. This is mostly because the models these wearables use are not trained for the movement patterns and EE of wheelchair users. Movement is different is different due to movement being constricted to the upper body, which also results in a lower resting EE [CB11].

Researching PA and EE for wheelchair users is challenging, because the models that are used to predict PA and EE that exist for non-wheelchair users do not work for wheelchair users. Their movement patterns are too different for a model that is not specially trained for wheelchair users to make accurate prediction. Self-reported measures, like questionnaires can be used, but they have been shown unreliable [She03]. For instance, activity intensity is very subjective for humans (light/heavy), compared to computers and sensors (kcal/min). Other measuring tools are either unreliable (current wearables) or unpractical (oxygen intake measuring masks or regular interviews). By building a model which can more accurately estimate the energy expenditure using only the data from a typical wearable, accessibility can be greatly improved.

1.1 EE estimation challenges and proposal

Besides improved feedback to wheelchair users, researchers conducting health studies could really benefit from an accurate algorithm for wheelchair users to estimate EE from just wrist-worn wearables, since it would prevent the need for the subjects to wear an expensive and uncomfortable oxygen intake measuring mask. EE data gives both researchers and health coaches a way to quantify the intensity of the activities, which in turn can be used to study how PA might help wheelchair users with their health problems. A recent example is Popp et al.'s study where EE and acceleration data was measured in 30 subjects during a multitude of physical activities, which was used to develop an EE estimation model [PRB+18]. Using the collected data and reference values from literature they were able to recommend daily physical activities for a healthy lifestyle to wheelchair-bound SCI patients. More accurate EE estimation methods would make health research for wheelchair users more accessible and more reliable.

More research in EE can also help wheelchair users with common diseases like obesity. Measuring EE is an important part of studies that help wheelchair users balance their calorie intake [FSGJ21]. Estimating EE has its challenges. Similar signals can have widely different EE, and widely different signals can have similar EE. Solving this is especially challenging for wheelchair users, because of their limited movement. Also, just the activity type, movement data and heart rate are not the only things that determine EE. EE is also highly dependent on the user's stamina/fitness and BMI. A higher BMI often correlates with a higher EE [PBCC96]. While statistics like height and BMI can just be asked from the user, stamina/fitness is a lot more subjective and harder to incorporate.

We can however use the acceleration and heart rate data to estimate the intensity of the activity. A higher intensity means a higher EE. While this can be very challenging, heart rate can give a decent indication of EE, but its relation is not perfect. The problem with having only accelerometery data for the movements is that in physics, *work* is the product of *distance* and *force* in that direction. Distance can be derived from measuring acceleration over time, but force can not. If someone were to pick up an empty bottle or a heavy metal object, the accelerometery data would be roughly the same. Only when you know what kind of activity the user is doing exactly, can you more accurately predict their EE. The goal of this thesis is to combine the activity type with acceleration and heart rate data which should give a regression algorithm better chances of accurately estimating EE.

The process of classifying the activity type is challenging, especially for wheelchair users. Many different signals can belong to one activity type. Let us take propulsion as an example. Some people will make a few long strokes, while others may make shorter strokes in quicker succession. The differences between users is also a challenge in itself. People could not only move their body differently because of different habits, but also because they might have different conditions, which is why this is especially a challenge for classifying the activity type for wheelchair users compared to non-wheelchair users. For example, someone with a lower limb amputation might sit less firmly/move their hips more freely which will lead to them having different movement patterns. Not only can different signals have the same activity type, but multiple activity types can also have the very similar signals. Since the accelerometery sensors are only on the wrists, the data only presents the movement of the wrist, not the whole body.

We propose a method to more accurately estimate EE for wheelchair users, where first the activity type is classified, which is then used to estimate the EE using regression. A supervised classification model is trained to predict the activity type. After adding the predicted activity to the other features, the specialised regression model for that specific activity can be used to estimate the energy expenditure. By building a model where the regression models are specialised in one specific activity type or class, we expect it to better interpret the accelerometery and heart rate data. Since it has more context on what the subject is doing, it should be able to perform better than a method without activity-specific EE estimation.

The dataset we used was provided from the DACT-Wheel (from Data to ACTion in Wheelchair users) project. During this project, a sample group was selected to partake in various physical activities, ranging from light to intense. The participants belonged to either of two groups: spinal cord injury (SCI) and lower limb amputation (LLA). During this project, the activity was also noted, which was used to develop a classification model. More information about the dataset is given in section 2. We evaluated the performance of our activity specific method and compared the results with those of a regular regression model, which are all based on random forests.

1.2 Background

Although research in EE estimation for wheelchair users is limited, there are many recent and ongoing studies to convert practices from non-wheelchair EE estimation to new methods for EE estimation in wheelchair users [NRTB17]. Activity-specific methods have also already been studied for non-wheelchair users. This will function as a starting point for this thesis. Albinani et al. tested multiple methods of using the predicted activity type to improve EE estimation, with their *MET lookup* method being the most effective [AIHR10]. MET stands for ‘metabolic equivalent’, with 1 MET being equal to the EE during a seated rest. During this study they experimented with different ways of using activity type detection to improve EE estimation. With one of their methods, instead of using some form of regression algorithm after having classified the activity type, they used a lookup table with MET values [AHW+00] for the activity values. After predicting the activity type, their method looks up the MET value for that activity and multiplies that with the EE for seated rest.

Albinali et al. noticed that with a perfect classifier, the MET lookup method would often underestimate the measured EE, which led to a MAPE of 88.9%. When they used their own classification method, which is not perfect, the MAPE would drop to 37.3%. They suspected that when a subject has a higher EE when doing an activity than usual, the classification method would classify it as a more intense activity, which made up for the underestimation of the MET table. One of the advantages of such a method is that it does not require EE measurements for training, like an oxygen measuring mask, which can generally only be used in lab settings.

For our method we decided on a different method to use activity types for EE estimation. A similar, well tested MET lookup table does unfortunately not exist for wheelchair users. Albinali et al. also touched upon using a method that was very similar to what we have done, were they used activity specific regression models to estimate EE, which was based on work by Crouter et al. [CKH+10]. For our method we decided upon random forest classification and regression models instead of C4.5 decision trees for classification and a linear, exponential or cubic regression model, which were used by Albinali et al. and Crouter et al. This was mainly for their wide range of application and ease of use.

It has been shown that current algorithms for EE estimation of wheelchair users are often far from adequate to support clinical research. Shwetar et al. [SVHD20] evaluated the performance of five modern algorithms and concluded that with MAPE (see section 2.5) scores ranging from 31% to 206% none of the algorithms were good enough. The goal of this research is to research if, and how, predicting the activity type would improve EE estimation.

Concerning the placements of the sensors for predicting EE of wheelchair users, studies have shown that a wrist-mounted accelerometer is the best choice. Nightingale et al [NWTB14] tested an ActiGraph accelerometer on the waist, upper arm and wrist of wheelchair users to predict EE during multiple activities. What they saw was that the EE predictions were most accurate when the accelerometer was placed on the wrist.

1.3 Thesis outline

Section 2 gives information about the data, preprocessing, features, labels and models that were used. The experiments and results are discussed in section 3. A conclusion is given in section 4, after which we further discuss the method and results in section 5, where also suggestions for future research are given.

2 Methods

In this section the workings of the activity-specific method are discussed in detail. First, information about the dataset that was used (section 2.1 and 2.2) will be given. Then we will explain how that dataset was preprocessed into features and labels (section 2.3) after which an explanation of what classification and regression models were used (section 2.4) and how their performance was measured (section 2.5) will be given. Lastly, an explanation on our proposed models is given (section 2.6). An overview of the activity-specific EE estimation method is given in Figure 1.

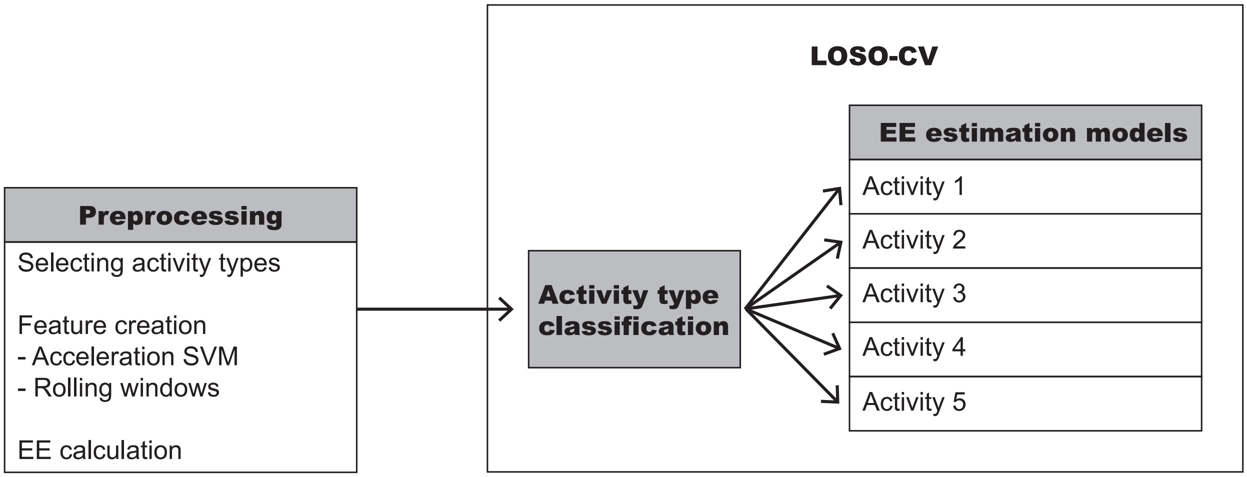


Figure 1: Schematic overview of the activity-specific EE estimation method.

The data used for this research was collected during the DACT-Wheel project, which is an ongoing project that started in 2017. The main goal of this project is to research potential positive effects that lifestyle and sports applications can have on wheelchair users.

The benefit of this data was that it was labelled and collected in a controlled environment, but it also has its downsides, because its generalisation with free living situations may be limited, which will be discussed in section 5. Several sensors were used, namely heart rate and accelerometery sensors on the wrists. This data would be similar to what a standard wearable, like a smartwatch, could produce, except for there being acceleration data from *both* wrists. Data was collected during exercises and daily activities, all while tracking the activity type.

2.1 Participants

The participants were wheelchair users with either a spinal cord injury (SCI) or a lower limb amputation (LLA). A spinal cord injury is damage to (part of) the spinal cord. Loss of muscle function and thus loss of limb control are symptoms which affects large parts of a patient's life. This often means that many SCI patients eventually end up in a wheelchair. Because people with SCI are limited in their daily movement, they often have a significantly lower energy expenditure [FG20]. As was described in the methods section of Hoevenaars et al. [HYH$^+$ew], the participants were recruited by advertisement via the Dutch SCI patients' association, social media and rehabilitation centre Reade in Amsterdam. All participants had to be between 18 and 75 years old, had SCI for at least one year, were not dependent on a ventilator and were wheelchair dependent for longer distances. If they had any of the following, they could not participate: presence of a pacemaker, severe edema, progressive illness, pressure ulcers, metabolic diseases, severe co-morbidities, psychiatric disorders, being pregnant, insufficient understanding of the Dutch language to understand the study. These characteristics were obtained by having the participants fill in a questionnaire and partake in an interview.

	SCI	**LLA**
Male/Female	26/13	6/6
Age mean	48.8	49.8
Age std	11.9	17.1
BMI mean	24.1	26.8
BMI std	4.6	9.0

Table 1: Information on the SCI and LLA participants

Extra information on the SCI and LLA groups are given in Table 1. The mean and standard deviation of age and BMI are given, together with the proportions of male and female subjects. The LLA group is considerately smaller, which results in less extensive training and a lower statistical power.

Two versions of the EE estimation model were built, which are identical in their workings, but use either the data from the SCI group or from the LLA group. Because of the differences in EE [EPS$^+$15, BP04], the data of these two groups were handled separately. In sections 3 and 5 the EE estimation performance differences are shown and discussed.

We discarded the measurements of subject where data was missing. Data could be missing because of various reasons, for instance one or more sensors were not working correctly, or when a patient did not complete all ADLs. After removing these subject from our dataset, there were 27 SCI patients and 10 LLA patients.

2.2 Sensors

The participants were fitted with a GENEActiv and an Activ8, which both contain accelerometers and were fitted on the left and right wrist respectively. A Fitbit containing a heart rate sensor was fitted on the wrist on the participant's non-dominant arm. A COSMED face mask which measures oxygen intake and carbon dioxide outtake was also fitted covering the mouth and nose. Other sensors were also used during the project, like an Activ8 on the right wheel of the wheelchair, but since our goal was to develop a model that could eventually be used for a wrist worn wearable, this data was not used. The locations of the sensors on the participants' body can be seen in Figure 2.

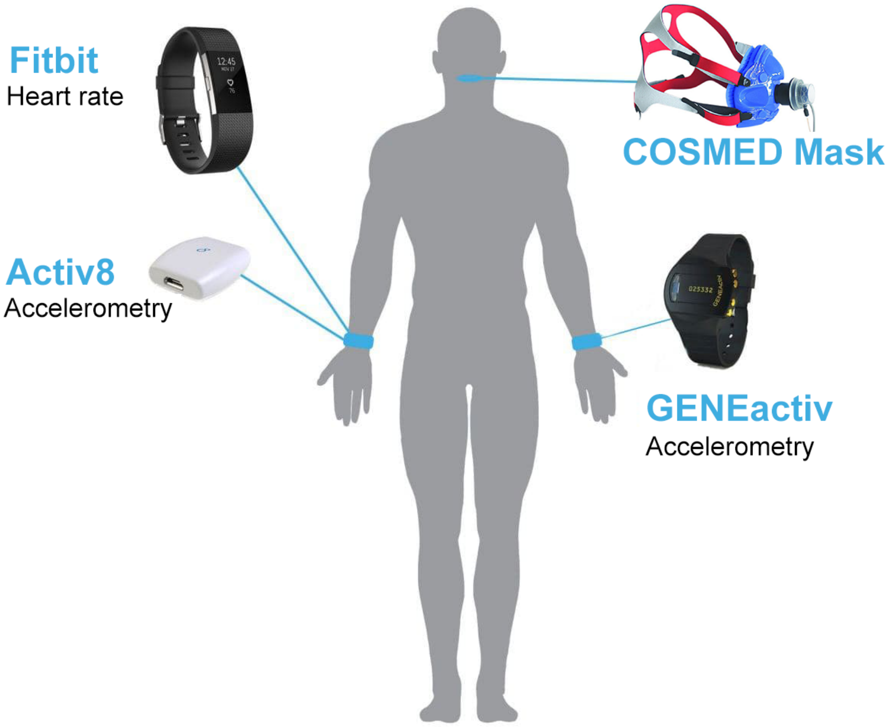


Figure 2: Locations of the sensors on the participants.

- **GENEActiv**
  The GENEActiv is a wrist-worn accelerometer designed by Activinsights. It outputs raw tri-axial accelerometery data, together with light and temperature measurements. It was designed to be used for health research and to be low burden for the user. It is lightweight, waterproof and ergonomic to make it easy to wear for long periods of time. Because it is a small, aesthetically neutral, wireless device that needs practically no maintenance besides charging every couple of days, it is excellent to use for monitoring everyday behaviours reliably [Act15].

During the DACT-Wheel project, the participants were equipped with a GENEActiv on their left wrist. It can continuously record accelerometery data at a sampling rate of 10 to 100 Hz and during the research, it was set to 100 Hz. It has been shown by third-party research to be reliable and give comparable results to other proven accelerometers. [PGCB16].

- **Activ8**
  The second accelerometer was attached to the right wrists of the participants. It also outputs raw tri-axial accelerometery, but at 12.5 Hz. While the GENEActiv measures continuously, the Activ8 records one second every three seconds (record 1 second, wait 2 second, repeat). These 'gaps' made the interpolation of this acceleration data extra challenging. The device was originally developed to be worn on the subject's thigh, but this was intentionally ignored, because the wrist and arm movements were deemed more relevant for this study, especially for wheelchair users.

- **Fitbit**
  The Fitbit would be classified as a 'sportwatch': a commercially available wearable with as main goal to help users with physical exercise. It is a consumer-grade wrist-worn device that was designed and marketed for consumer use, but its heart rate data has shown reliable enough to be used for health research [HYH+ew]. The device continuously measures the changes in blood pressure, which correspond with blood volume changes. An internal algorithm can estimate the actual heart rate using this data, which it then logs every 5 to 15 seconds. The Fitbit can also record and estimate multiple other variables, but for this research we only used the heart rate data.

- **COSMED Mask**
  COSMED provides face masks for measuring the amount of oxygen a person consumes and the amount of carbon dioxide they exhale. The energy expenditure can be derived from the oxygen intake and carbon dioxide outtake using Weir's formula [Wei49]. The face mask is designed to be comfortable for multiple hours and at both light and heavy exercise intensities, while creating an airtight seal around the mouth and nose to prevent any leaking and false measurements.

Sensor	**SCI**	**LLA**	**Total**
GENEActiv	1459361	569026	2028387
Activ8	400575	164814	565389
Fitbit	3048	1362	4410
COSMED	7132	2908	10040

Table 2: Number of measurements per sensor.

As is shown in Table 2, the amount of measurements per sensor is very different. This is partially because the devices were not activated and deactivated at the same time when the research was conducted, but the main reason for the differences in number of samples is because of the sampling rate. The GENEActiv has a sampling rate of 100 Hz, while the Activ8 measures at 12.5 Hz [FG20]. The Fitbit logs the heart rate every 5-15 seconds [HYH+ew]. Using the timestamps, the acceleration data (GENEActiv and Activ8) could be down-sampled and up-sampled respectively to 32 Hz (every

0.03125 seconds), to make it easier to work with. While the GENEActiv measures continuously, the Activ8 measures one second every three seconds (there are 32 samples in one second, then two seconds nothing, repeat). The COSMED mask measures at an unfixed sampling rate, based on breath rate, with on average one sample per 6 seconds. These numbers also show that there will be very few lines in the dataset were all sensors created a datapoint all at the same time. The next section gives an explains how this issue was solved.

2.3 Preprocessing

The aforementioned sensors produce both raw data and processed information. For instance, the Activ8 has an internal algorithm to classify movement and estimate EE for healthy users. Since it was worn on the wrists by wheelchair users instead of the thigh by healthy users, this information is not used. For this research we developed features from the raw data.

GEN_acc.x **GEN_acc.y** **GEN_acc.z**	Tri-axial accelerometery from the GENEActiv on the left wrist
ACT8RWrXYZ_x **ACT8RWrXYZ_y** **ACT8RWrXYZ_z**	Tri-axial accelerometery from the Activ8 on the right wrist
FBsec_value	Heart rate from the FitBit on the right wrist
GEN_Activity	The activity type
COS_V.O2	The measured oxygen intake from the COSMED mask
COS_V.CO2	The measured carbon dioxide outtake from the COSMED mask

Table 3: All selected data to build features and labels.

The GENEActiv accelerometery was downsampled to 32 Hz and the Activ8 accelerometery had to be upsampled. The 'gaps' in the Activ8 measurements (as described in section 2.2) were not interpolated, so they were still present in the data.

2.3.1 Features

Firstly, we also calculated the signal vector magnitude (SVM) for both the accelerometers with the formula $\sqrt{x^2 + y^2 + z^2}$. The vector magnitude is the length of the vector, or total accelerations of the sensors. This resulted in four data-points per accelerometer: x, y, z and SVM. Because of earth's pull there will always be a downward component, but since the orientation of the accelerometers is not fixed, this cannot simply be corrected for without knowing the orientation of the accelerometer, which is constantly changing. Just subtracting earth's gravitational acceleration from the downwards acceleration is only possible if you know what way downwards is. In other words, if the vector magnitude is equal to zero, the accelerometer is accelerating towards earth at the same rate as the gravitational acceleration, also known as 'falling'. If the vector magnitude is exactly earth's gravitational acceleration, the accelerometer could be stationary, but also accelerating horizontally while falling, accelerating downwards at twice earth's pull or a combination of both. Because we use random forests for our classification and regression models the features do not need to be normalised to values between 0 and 1. The reason for this is explained in section 2.4. There are methods to subtract the gravitational acceleration from the measured signal [VHGDL+13], but these were not used in this research.

Just using the raw acceleration data as features would not work, since that would only give an indication of what direction and speed the participant's arm is accelerating at that single moment. Features relating to the acceleration data would optimally capture the exact patterns of the movements. This pattern would then relate to the type of activity, so a classification model can be trained on those features.

Another challenge is both the uneven sampling rates between the sensors and the gaps in the data. The classification and regression models from scikit-learn only work when no values are missing, but due to the different sampling rates the data rarely lines up. This means that any samples with missing values either have to be filtered out, be shifted to line up or be imputed with new values by interpolation.

Both challenges can be solved by using a moving average (also known as a rolling mean), which for every sample calculates the mean acceleration of the previous n samples. This way, we get information about the movement of a timeframe instead of one moment. If the window size is big enough to cover the gaps in the data, it can be used to interpolate the data from sensors with a lower sampling rate. This can all be implemented using the `rolling` function from the Pandas library. This function allows us to perform calculations on rolling windows. Using this, we can set a window size and calculate the moving average, which for every sample calculates the mean acceleration of the last n samples. If n is big enough to cover the gaps in the data, which are rarely bigger than 2 seconds (64 samples at 32 Hz), the created feature will continuously fill every sample with a value. Any missing samples will be ignored when calculating the moving average by setting `min_periods` to 1. The size of the rolling window was set to 400 samples for acceleration and 150 samples for heart rate (12.5 seconds and 4.7 seconds at 32 Hz). This means that for all features relating to acceleration, calculation are done from all previous samples that are no more than 12.5 seconds old. Likewise for heart rate features with samples of no more than 4.7 seconds old. These separate window sizes were required through optimisation.

Using the same rolling window technique, we can also calculate the moving median, standard deviation, minimum and maximum values of the acceleration and heart rate. These statistics give a decent indication of the movement and heart rate trend of the user and were used as features.

A complete list of all features can be seen in Table 7. As you can see, the rolling standard deviation for the heart rate was left out. This was done because for the window size we used, there were not enough values to calculate a meaningful standard deviation and Panda's `std` function was undefined.

Figure 3 shows a correlation matrix of all the features that were used. It clearly shows a high correlation between the minimum/maximum values and the standard deviation. This makes sense because a high standard deviation would naturally lead to more extreme minimums and maximums. What's interesting is that the acceleration from the left and right wrist are also correlated, although not as extreme. Also, the correlation of the means and median of the different sensors can be seen through the diagonal lines in the upper left corner of the correlation matrix. Lastly, the thing that stands out through the red colouring is the negative correlation between the minimum values and the other values, especially maximums and standard deviations of the accelerometer data. This means that a low minimum will often be paired with a high maximum and deviation, which is exactly what happens during quick or extreme movements.

As can also seen in Figure 3, the maximum, minimum and standard deviation of the acceleration and heart rate features are all correlated with EE, just like the acceleration magnitudes of both sensors. This is a promising sign for their use as features for EE estimation. If a feature is not directly correlated with EE, that would not always mean that feature should be left out for EE estimation. One feature might for instance only show a high correlation with EE when paired with other features.

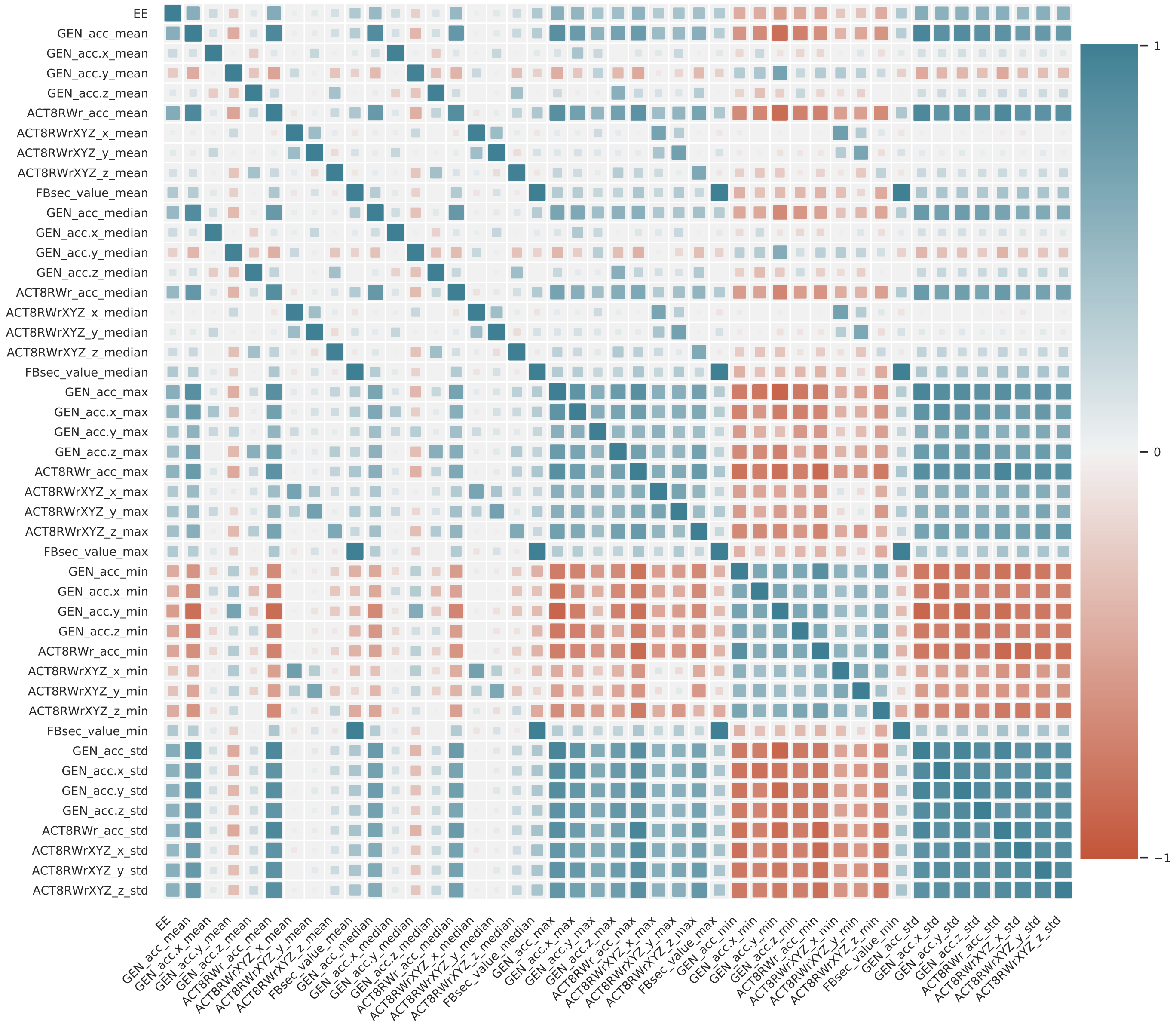


Figure 3: Correlation matrix of the features and EE.

2.3.2 Activities

The activity type can be seen as both the label for the classification model and a feature for the regression model. During the DACT-Wheel project, the participants did both activities of daily living (ADL) and strength exercises. Since the type of strength exercise varied from person to person, we only used the data for the ADLs and the seated rest.

ADL	Description
1	Move the wheelchair forward at slow speed for 60 seconds.
2	Move the wheelchair forward at medium speed for 60 seconds.
3	Move the wheelchair forward at high speed for 60 seconds.
4	Use of a hand ergometer for 60 seconds.
5	Handing over objects from a bag to the researcher one by one, while the wheelchair is propelled by the researcher for 60 seconds. (Left hand and right hand alternately)
6	Simulate the process of setting a table with plastic cups and plates for 60 seconds.
7	Simulate the process of doing the dishes for 60 seconds.
8	Simulate the process of using a laptop positioned on a table for 60 seconds.
9	Perform slaloms around 5 pieces that are 1,5 meters apart from each other for 60 seconds.
10	Simulate playing wheelchair basketball for 60 seconds. (Rolling, bouncing and throwing the basketball)
11	Perform a transfer from the wheelchair to a bed and vice versa for 60 seconds.

Table 4: Daily activities performed by the sample group.

Another option is to divide these ADLs into two activity classes to reduce the number of possible labels and increase simplicity. As long as the activities in the same class are similar, the classification model will be more accurate. The regression models will be less specialised, but if the relation between movement patterns and EE within one class are similar, this would not be problematic. One way to do this is by looking at activities with or without propulsion. As you can see in Table 5, ADL 6, setting the table, is a combination of propulsion and no propulsion. We decided to classify this as if it were propulsion, since this showed the best results during our testing.

ADL	Activity	Activity class
1	Propulsion: slow	Propulsion
2	Propulsion: normal	Propulsion
3	Propulsion: fast	Propulsion
4	Handcycling	No propulsion
5	Rummaging in bag	No propulsion
6	Setting table	Combination
7	Doing dishes	No propulsion
8	Typing on laptop	No propulsion
9	Manoeuvring wheelchair	Propulsion
10	Basketball	Propulsion
11	Transfer to chair	No propulsion

Table 5: Activity class per activity type.

Table 6 shows some of the differences between EE and acceleration data between the twelve activity types for the SCI group. The mean and standard deviation of both the EE and acceleration magnitude of the GENEActive sensors are given. As can be see the mean EE per activity type differs a lot, and is clearly higher on more intense activities. The standard deviation for the acceleration is lowest for a seated rest. A differences in EE and acceleration can be most easily recognised when comparing slow, normal and fast propulsion (ADL1, ADL2 and ADL3). This shows how a classification model can recognise activities and how a EE estimation method can benefit from activity-specific approaches.

GEN_Activity	**GEN_acc_mean mean (std)**	**EE mean (std)**
rest	1.011 (0.015)	1.853 (0.721)
adl1	1.080 (0.064)	2.066 (0.651)
adl2	1.178 (0.125)	2.422 (0.851)
adl3	1.412 (0.278)	3.602 (2.693)
adl4	1.147 (0.113)	2.289 (0.619)
adl5	1.052 (0.047)	1.970 (0.549)
adl6	1.036 (0.032)	2.124 (0.715)
adl7	1.043 (0.020)	2.037 (0.513)
adl8	1.022 (0.020)	1.979 (0.346)
adl9	1.162 (0.092)	2.507 (1.313)
adl10	1.194 (0.072)	2.695 (1.105)
adl11	1.075 (0.045)	3.022 (1.419)

Table 6: Statistics on the differences in EE and acceleration properties per activity type for the SCI group.

2.3.3 Target

The energy expenditure was calculated from the data from the COSMED mask using Weir's formula [Wei49]:

$$\mathrm{EE} = 3.942 \cdot \mathrm{VO2} + 1.106 \cdot \mathrm{VCO2}$$

VO2 and VCO2 are the oxygen intake and carbon dioxide outtake respectively, both in litres per minute. The result is the energy expenditure in kilo-calories per minute. The sampling rate of the COSMED mask is much lower than that of the accelerometers, since it is linked to breath rate, but this means the data file contained mostly empty values. Naturally, EE could only be calculated for the non-empty values. The sampling frequency was not changed during this progress, so the target sampling rate is equal to that of the COSMED mask, which is dependent on the breath rate, averaging at one sample per 6 seconds.

Summarised, the preprocessing went as follows: first the data from unused activities were filtered out. Then rolling windows were applied on the acceleration and heart rate data, which also filled all gaps, so there were no empty values in the features. EE was then calculated for every sample of the COSMED mask. All rows without an EE value were removed and the end result was a table of features and labels without any empty values.

	Rolling mean	**Rolling median**	**Rolling maximum**	**Rolling minimum**	**Rolling standard deviation**
GENEActiv magnitude	GEN_acc_mean	GEN_acc_median	GEN_acc_max	GEN_acc_min	GEN_acc_std
GENEActiv x-axis	GEN_acc.x_mean	GEN_acc.x_median	GEN_acc.x_max	GEN_acc.x_min	GEN_acc.x_std
GENEActiv y-axis	GEN_acc.y_mean	GEN_acc.y_median	GEN_acc.y_max	GEN_acc.y_min	GEN_acc.y_std
GENEActiv z-axis	GEN_acc.z_mean	GEN_acc.z_median	GEN_acc.z_max	GEN_acc.z_min	GEN_acc.z_std
Activ8 magnitude	ACT8RWr_acc_mean	ACT8RWr_acc_median	ACT8RWr_acc_max	ACT8RWr_acc_min	ACT8RWr_acc_std
Activ8 x-axis	ACT8RWrXYZ_x_mean	ACT8RWrXYZ_x_median	ACT8RWrXYZ_x_max	ACT8RWrXYZ_x_min	ACT8RWrXYZ_x_std
Activ8 y-axis	ACT8RWrXYZ_y_mean	ACT8RWrXYZ_y_median	ACT8RWrXYZ_y_max	ACT8RWrXYZ_y_min	ACT8RWrXYZ_y_std
Activ8 z-axis	ACT8RWrXYZ_z_mean	ACT8RWrXYZ_z_median	ACT8RWrXYZ_z_max	ACT8RWrXYZ_z_min	ACT8RWrXYZ_z_std
Fitbit heart rate	FBsec_value_mean	FBsec_value_median	FBsec_value_max	FBsec_value_min	

Table 7: All features that were used by the random forest models.

2.4 Models

Activity type classification was done using a random forest classification model. Multiple random forest regression models were used for the energy expenditure regression. These models combine the advantages of using decision tree algorithms and bagging algorithms. The descriptions of these methods are based on Witten's book on data mining [WFH11].

- **Decision trees**
  One of the most popular predictive models are decision trees, mainly because of their simplicity. It operates by labelling every non-leaf node with an input feature and labelling the branches with a 'decision' based on those input features. The leaves contain the category (with classification) or a continuous value (with regression). Training is done by calculating the information entropy before and after splitting on an input feature. The difference between these entropies is the information gain and the higher the information gain, the better that split is. Decision trees do not require data normalisation, because of the way the trees are trained. By solely looking at information gain and entropy, the range of the values does not matter. Decision trees are unfortunately non-robust, prone to overfitting and are biased to the majority class when working with imbalanced classes [LCCC10]. The first two problems can be partially solved by pruning setting a maximum depth, which both reduce the complexity of the decision tree.

- **Bagging**
  Bagging, which is short for bootstrap aggregating, is a ensemble meta-algorithm to improve the performance of a black box estimator (a prediction model). It works by drawing random samples with replacement which results in multiple sets of the same size as the original training set. This part is called bootstrapping. Multiple instances of the black box estimator are trained on these sets and their result is either the most common category (with classification) or the mean output of estimators (with regression). This technique makes it possible to 'combine strengths' of multiple weak learners and make it perform better than a single learner would. The technique does not alter the way the individual learners operate, so overfitting is still a challenge when working with decision trees, but by aggregating those learners, this can be significantly improved. Lastly, having to train multiple instances is naturally a lot more computationally expensive, but because the individual learners do not interact with each other in any way, training and prediction can be performed in parallel.

- **Random forests**
  Random forests are an ensemble of multiple decision trees and it is basically a combination of the techniques used in decision trees and bagging. By introducing randomness in the training process, diversity can be achieved among multiple decision trees. This randomness is mainly implemented in two ways. The first random factor is that the training set for each decision tree in the forest is a random sample (with replacement) from the total training set (bagging). By randomising the subspace, only a subset of the features are used by each decision tree. This is also known as 'feature bagging'. The second factor is that the decision tree training procedure is modified to be random instead of deterministic.

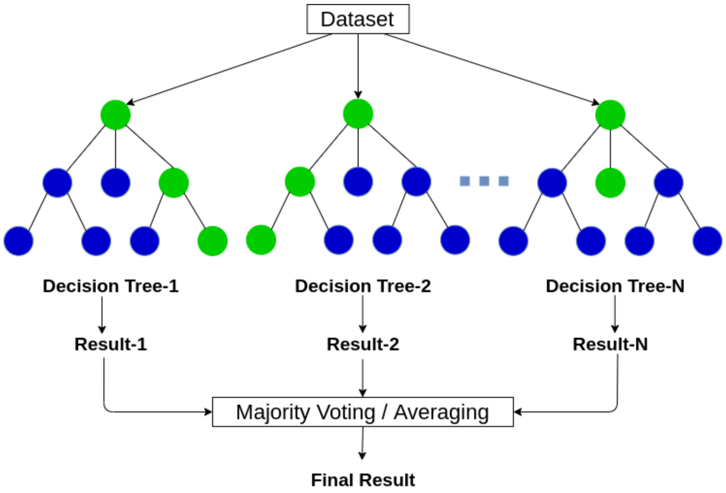


Figure 4: Diagram of a random forest model. Source: www.analyticsvidhya.com

> Random forests work as follows: Multiple decision trees are constructed and during testing/predicting, each tree will give a predicted value. In case of classification, this will be a class and the most predicted class will be the output of the random forest classification model. This is known as a majority vote, because the class with the majority of the votes will be selected as the output. In case of regression, the trees will output a continuous value which can be averaged. The main benefit of using random forests over decision trees is that although decision trees often tend to overfit, training multiple randomised decision trees on randomised data, the average of their outputs will be less likely to overfit. A diagram of a random forest can be seen in Figure 4.

For our implementation, scikit-learn's default hyperparameters were used for all random forests, besides the maximum depth being set to 12 (instead of no maximum) to reduce overfitting and the number of trees being upped from 100 to 200 for a slight performance increase. During experimentation, other hyperparameters were tested, but this showed no significant performance increase.

2.5 Performance evaluation

Training and testing on the same dataset should never be done, because a model that performs well on the data it has trained on often will perform worse when presented with new data. To prevent this overfitting problem, we needed to split the dataset into independent subsets. For the training and test set to be independent the training and test group would ideally include different subjects to prevent bias [SLJ+17]. If data from one subject is in both the training and test set, a model may be overfitted to this subject's data, which would make it perform well on the test set. If it was then presented with data that is actually new and independent data, the overfitted model will perform worse than on the test set. For this reason, we used a subject-wise cross-validation method known as *leave-one-subject-out* cross-validation (LOSO or LOSO-CV) to make sure the training and test set contained no mutual subjects so they are independent.

The simplest and most common cross-validation (CV) methods are k-fold CV and leave-one-out CV. With k-fold CV, the original dataset is split into k subsets. One of those subsets will be the test set and the other $k-1$ sets will be the test set. This process is then repeated k times until every subset has been used as test set exactly once.

LOSO works by taking a single subject (participant) as the test set and all other subjects as the training set, which is then repeated for every subject. In this way, the training and test sets will never include samples from the same subject. We implemented LOSO using the `LeaveOneGroupOut` function from the scikit-learn Python library. In our implementation, every sample from one subject was assigned to the same group.

To evaluate the performance of a regression model, you need to compare the predicted labels with the labels in the test set. For a regression model, this can be done using the mean absolute percentage error (MAPE):

$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

A_t is the actual value and F_t is the predicted (forecast) value. MAPE can be calculated using the `mean_absolute_percentage_error` function from the scikit-learn library. This function returns a floating point between 0 and 1, so not as an actual percentage.

The performance of the classification model can be evaluated with *accuracy*, *precision* and *recall*:

$$\begin{aligned}
\text{accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\
\text{precision} &= \frac{TP}{TP + FP} \\
\text{recall} &= \frac{TP}{TP + FN}
\end{aligned}$$

$$\begin{aligned}
TP &= \text{True Positive} \\
TN &= \text{True Negative} \\
FP &= \text{False Positive} \\
FN &= \text{False Negative}
\end{aligned}$$

To evaluate the performance of our models, we calculated the performance for every fold, or subject. This was then all added up and divided by the number of subjects to give a total score/error. For example, the accuracy scores that are mentioned in the methods sections are actually the mean average over all folds. For simplification, these evaluation methods will just be named by their usual names.

2.6 Model comparison

To be able to measure any potential improvements that our new method would bring, a basic random forest regression model was built to assess the expected performance in estimating EE. This model only uses the features (Figure 7) and EE as the target. The activity type was not used for the baseline model. The dataset was split using LOSO-CV and evaluated using MAPE (see section 2.5).

Just like we can use a random forest model for regression, there is also a version available in the scikit-learn library for classification. We can modify the regression model for activity type classification by changing the model to classification, switching the label from EE to the activity type and changing the performance metrics. The resulting model will classify the activity instead of estimating EE. If we use all ADLs as labels, there are 12 labels in total (11 ADLs and rest), and when we divide the ADLs into two classes, there are 3 (propulsion, no propulsion and rest).

After training the activity type classification model, we had to find a way to use the predicted activity to improve the EE estimation process. We ended up with two versions; one with twelve regression models and one with three. After splitting the dataset using LOSO cross validation, the classification model is trained using the training set and then predicts the activity type of all samples in the test set. The training set is then split into subsets based on the *actual* activity types. These subsets are used to train each of their respective regression models. The test set is then split, but based on the *predicted* activity types. EE is then estimated for all of these test subsets, also by their respective regression models. This whole progress is then repeated until the LOSO algorithm is done. The result is essentially a list of either three or twelve regression models, each trained for their specific activity type, where a classification model chooses which regression model to use to predict EE.

The training of the regression models can be done in two ways. The first option is to split the training set into subsets using the actual activity type. This way, each regression model will be trained completely for a specific activity. The second option is to first train and run the classification model and use the activities that it predicts to split the training set on. Training on the actual activity type will make the individual regression models more specialised for their activity. Training on the predicted activity, on the other hand, may lead to the regression models being trained for what the classification model will predict the activity to be e.g. normal propulsion might often mistakenly be classified as fast propulsion, which might cause the regression model to better at predicting these cases and less over-estimations. We tested both options in section 3.

3 Results

To understand if and how our proposed method had any effect on the performance of EE estimation, we first fitted a baseline model to assess baseline performance. To then be able to assess the performance and limitations of activity-specific EE estimation, the performance of both the classification model and regression model were evaluated, to see what parts of the activity-specific EE estimation could be limiting the overall performance. In summary, the performance of the baseline method and the activity-specific method (both classification and regressions parts) will all be discussed in this section.

We experimented with both types of labels for the classification model (section 2.3.2). The first variant, where the twelve activity types are used, consists of twelve regression models. The second variant uses the three activity classes, so there are only three regression models, and thus more training (and testing) samples per model. As was mentioned in section 2.1, the models can use either the data from the SCI patients or from the LLA patients.

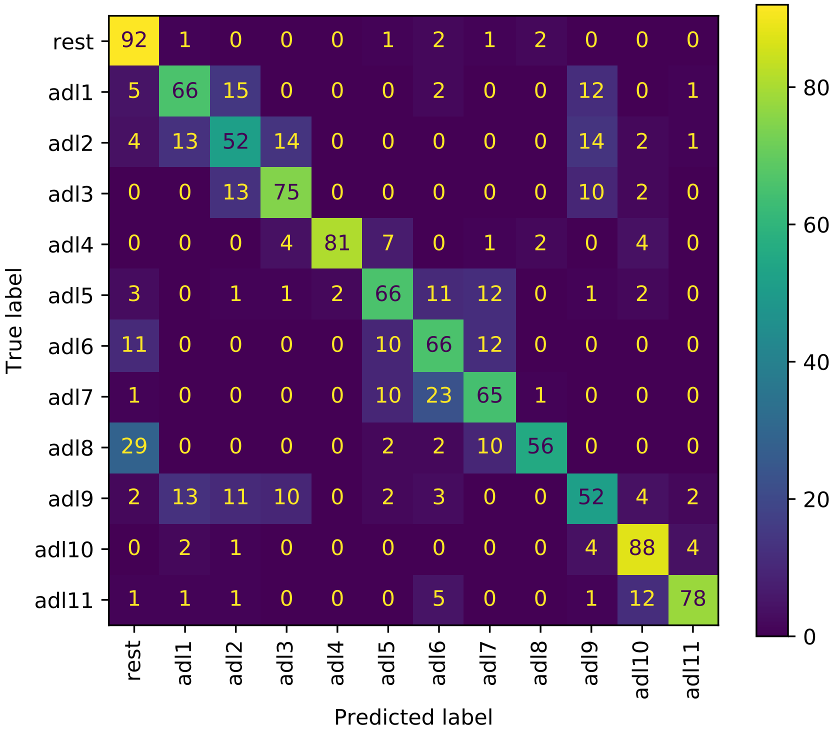
3.1 Classification

The results from the activity type classification model for the SCI and LLA groups will be evaluated in the next two subsections. Because of the differences in EE (see section 2.1 and Table 1), the data of these two groups were handled separately.

3.1.1 Spinal cord injury classification

When classifying the twelve activity types, the SCI classification model has an accuracy of 0.732. By noting the predicted activities alongside the actual activities, we can construct a row-normalised confusion matrix (Figure 5). A confusion matrix allows as to visualise and interpret the performance of the classification model. The values are normalised on the actual labels (rows), so the numbers in any field are the fraction of samples in that actual activity (row) that were classified as that predicted label (column). The numbers are presented as percentages to improve readability. As a result, the values on the diagonal are equal to the recall scores. As can be seen in the confusion matrix depicted in Figure 5, the mistakes were made on the *rest* label, with 91% being recognised as such by the random forest classification model.

It is interesting to see how for instance the classification model often mistakes normal propulsion (ADL2) for slow propulsion, fast propulsion and slalom (ADL1, ADL3 and ADL9). This is most likely because the movement patterns for these activities are similar and are all propulsion (see Table 4 and 5). The confusion matrix also clearly shows the distribution of the classes. While *rest* is a lot more common, all other classes are well balanced. Since tree-based methods are usually biased towards the majority class, it is important to pay extra attention to the precision scores of the majority class. A low precision of the majority class could mean that the model is biased to (falsely) classifying any activity type as the majority class. Despite this class imbalance and tree-based methods usually being biased towards the majority class, which is *rest* in this case, the precision for the majority class is as high as 0.85. A high precision for *rest* means that few other classes are mistakenly classified as *rest*.



	rest	adl1	adl2	adl3	adl4	adl5	adl6	adl7	adl8	adl9	adl10	adl11
rest	92	1	0	0	0	1	2	1	2	0	0	0
adl1	5	66	15	0	0	0	2	0	0	12	0	1
adl2	4	13	52	14	0	0	0	0	0	14	2	1
adl3	0	0	13	75	0	0	0	0	0	10	2	0
adl4	0	0	0	4	81	7	0	1	2	0	4	0
adl5	3	0	1	1	2	66	11	12	0	1	2	0
adl6	11	0	0	0	0	10	66	12	0	0	0	0
adl7	1	0	0	0	0	10	23	65	1	0	0	0
adl8	29	0	0	0	0	2	2	10	56	0	0	0
adl9	2	13	11	10	0	2	3	0	0	52	4	2
adl10	0	2	1	0	0	0	0	0	0	4	88	4
adl11	1	1	1	0	0	0	5	0	0	1	12	78

Figure 5: Confusion matrix of the SCI classification model with all 12 activity types.

From the results of such a classification model, the precision and recall score for every activity type can be calculated:

Activity	**Precision**	**Recall**
Rest	0.85	0.91
ADL1	0.64	0.67
ADL2	0.59	0.53
ADL3	0.76	0.76
ADL4	0.94	0.83
ADL5	0.69	0.66
ADL6	0.55	0.69
ADL7	0.65	0.64
ADL8	0.79	0.55
ADL9	0.59	0.56
ADL10	0.83	0.88
ADL11	0.83	0.79

Table 8: The precision and recall scores with all 12 activity types for the SCI model.

Scaling down the number of labels leads to a higher accuracy, but less specialised regression models, so it is a trade-off. When the number of labels are scaled down from twelve to three by using activity classes instead of activity types as described in section 2.6, the accuracy improves to 0.882, which means that significantly less mistakes are made (−44%). As can be seen by the confusion matrix depicted in Figure 6, the classification model sometimes confuses *propulsion* with *no propulsion*, but the overall fraction of false predictions is much lower. Just like with the twelve activity type classification, the three classes are imbalanced, with *propulsion* more common than the other classes. (1087 *rest*, 2392 *propulsion* and 1452 *no propulsion* in SCI group). While this is usually problematic for decision tree based models, the precision for the *propulsion* class is 0.91, which is the highest of the three classes (see Table 9).

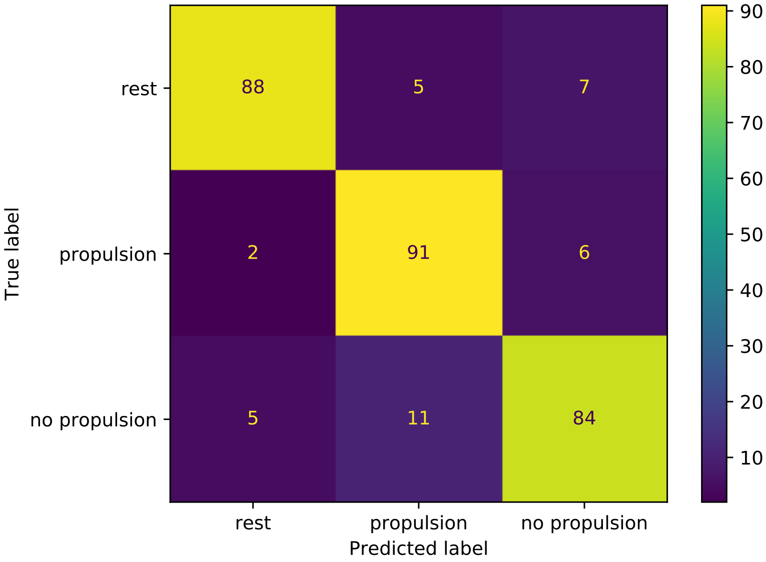


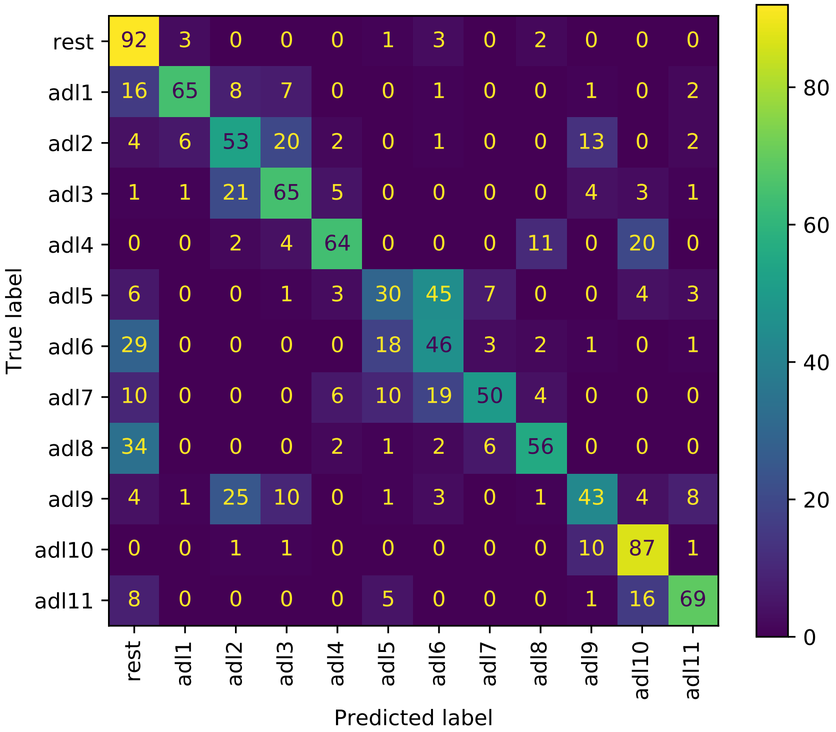
Figure 6: Confusion matrix with the 3 activity classes for the SCI model.

Activity	**Precision**	**Recall**
Rest	0.89	0.87
Propulsion	0.91	0.91
No propulsion	0.84	0.85

Table 9: The precision and recall scores with the 3 activity classes for the SCI model.

3.1.2 Lower limb amputation classification

When we repeat this process with the model for the LLA subjects, we get an accuracy of 0.660 when classifying the twelve activity types. We can again construct a confusion matrix (Figure 7) and a table of precision and recall scores (Table 10) to analyse the performance.



	rest	adl1	adl2	adl3	adl4	adl5	adl6	adl7	adl8	adl9	adl10	adl11
rest	92	3	0	0	0	1	3	0	2	0	0	0
adl1	16	65	8	7	0	0	1	0	0	1	0	2
adl2	4	6	53	20	2	0	1	0	0	13	0	2
adl3	1	1	21	65	5	0	0	0	0	4	3	1
adl4	0	0	2	4	64	0	0	0	11	0	20	0
adl5	6	0	0	1	3	30	45	7	0	0	4	3
adl6	29	0	0	0	0	18	46	3	2	1	0	1
adl7	10	0	0	0	6	10	19	50	4	0	0	0
adl8	34	0	0	0	2	1	2	6	56	0	0	0
adl9	4	1	25	10	0	1	3	0	1	43	4	8
adl10	0	0	1	1	0	0	0	0	0	10	87	1
adl11	8	0	0	0	0	5	0	0	0	1	16	69

Figure 7: Confusion matrix with the 12 activity types for the LLA model.

Notice how less of the samples are on the diagonal, which means the LLA classification model more often predicted labels differently than their actual class compared with the SCI model. This results in a lower accuracy score (0.660 versus 0.732) and recall scores as low as 0.30. ADL5 is more often mistaken for ADL6 than correctly predicted. It is difficult to say with full certainty what the exact reason for the lower accuracy is, but a much smaller training set will generally result in worse performance.

Activity	**Precision**	**Recall**
Rest	0.76	0.92
ADL1	0.78	0.65
ADL2	0.51	0.53
ADL3	0.61	0.65
ADL4	0.74	0.64
ADL5	0.42	0.30
ADL6	0.38	0.46
ADL7	0.77	0.50
ADL8	0.70	0.56
ADL9	0.56	0.43
ADL10	0.73	0.87
ADL11	0.75	0.69

Table 10: The precision and recall scores with all 12 activity types for the LLA model.

When we use the three activity classes as labels, the accuracy increases to 0.810, which follows the same pattern as the SCI classification model. The LLA model still does not perform as well as the SCI model (accuracy = 0.882), most likely for the same reasons as when using the twelve activity types as labels. The confusion matrix can be seen in Figure 8 and the precision and recall scores in Table 11.

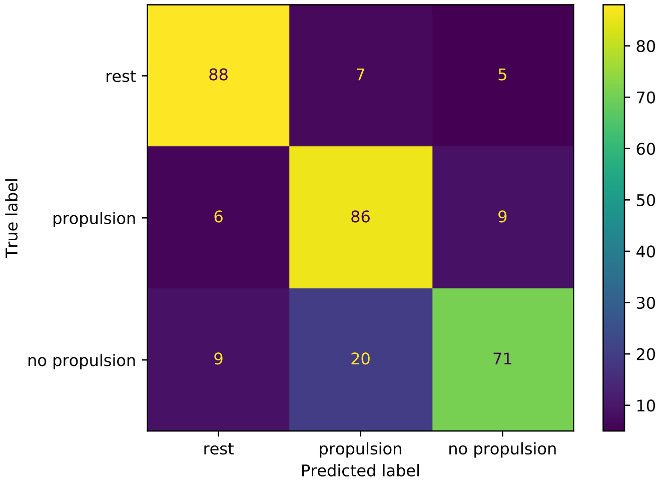


Figure 8: Confusion matrix with the 3 activity classes for the LLA model.

Activity	Precision	Recall
Rest	0.81	0.88
Propulsion	0.84	0.86
No propulsion	0.80	0.71

Table 11: The precision and recall scores with the 3 activity classes for the LLA model.

3.2 Energy expenditure estimation

As described in section 2.6, a baseline score is required to be able to understand if and how the activity-specific estimation method would result in improved EE estimation. When the baseline regression model is ran, which does not make use of the activity type data and is a basic random forest regression model, it gets a mean MAPE score of 0.367 for the SCI group and 0.336 for the LLA group. The target frequency is equal to that of the COSMED mask, which is unfixed and averages to 6 seconds per sample.

In this thesis we proposed four variants of using the activity type for the activity-specific EE estimation. Firstly, you can either use the *actual* activity types to train the individual random forest regression models for every activity, or the activity types that were *predicted* by the classification model (see section 2.6). Secondly, you can either classify the twelve activity types or the three activity classes. More classes means less accurate classification, but more specialised regression. The MAPE scores for each of these four variants are noted in Table 12.

Trained using	MAPE
predicted 12 types	0.368
actual 12 types	0.360
predicted 3 classes	0.358
actual 3 classes	0.350
Baseline regression	0.367

Table 12: MAPE for ways of training the SCI regression models.

As can be seen in Table 12, using the actual activity instead of the predicted activity to train the models in both cases results in a slightly lower error. Using the three classes instead of all twelve activity types, lowers the MAPE a bit further. This can partially be because the classification algorithm makes significantly fewer mistakes (accuracy increase from 0.732 to 0.882) and because activity types within the same class are indeed very similar. Also, more samples per class often means the models can be trained to perform better.

By analysing the individual MAPE scores per fold using a Kolmogorov–Smirnov test, we can see that the data does not differ significantly from a normal distribution ($p = 0.35$ for baseline regression). For this reason, we can use a paired t-test the see if the different methods are significantly better. A paired t-test can be used to assess of population means differ with two samples, where the groups are related. Each fold represents a subject, and we performed LOSO-CV on the same set of subjects for every method.

The t-values are noted in Table 13 and significant t-values ($\alpha = 0.05$) are marked. Since the differences in performance are small, only the model that was trained on the actual three activity classes performed significantly better than the baseline regression model. The difference in performance of the other models compared with the baseline model is so small, that it is too likely that the performance differences are just by chance.

	Baseline	*predicted* 12 types	*actual* 12 types	*predicted* 3 classes	*actual* 3 classes
Baseline	x				
predicted 12 types	0.4032	x			
actual 12 types	0.2849	1.3907	x		
predicted 3 classes	1.7536	2.1161*	0.9964	x	
actual 3 classes	3.1312*	3.0499*	1.9284	2.5692*	x

Table 13: t-values for paired t-tests on the MAPE scores of all folds of the SCI group.

When we repeat these experiments for the LLA group of the activity-specific EE estimation method, we get the results that are noted in Table 14. As can be seen, none of our methods improved over the baseline method and when we again perform paired t-tests on the MAPE scores of the individual folds (Table 15), the differences between the performance where all so small, that there were no significant differences at all.

Trained using	**MAPE**
predicted 12 types	0.346
actual 12 types	0.350
predicted 3 classes	0.347
actual 3 classes	0.343
Baseline regression	0.336

Table 14: MAPE for ways of training the LlA regression models.

	Baseline	*predicted* 12 types	*actual* 12 types	*predicted* 3 classes	*actual* 3 classes
Baseline	x				
predicted 12 types	0.2647	x			
actual 12 types	0.3729	0.3602	x		
predicted 3 classes	0.4821	0.0350	0.1977	x	
actual 3 classes	0.3589	0.1567	0.3707	0.7368	x

Table 15: t-values for paired t-tests on the MAPE scores of all folds of the LLA group.

4 Conclusions

In this thesis, we used an activity-specific EE estimation method specifically for wheelchair users. With this method the knowledge of the activity type was combined with the acceleration and heart rate data to estimate EE. The baseline model that used a basic random forest regression model resulted in a MAPE af 0.367 when using the SCI group. Using twelve activity types gave a result that was not statistically significant enough to consider it an improvement over the baseline model. Using three regression models that were trained on the actual activity classes gave the best result with a MAPE of 0.350. When using the LLA group, the activity-specific method did not beat the baseline method and the results were not statistically significant.

The accuracy of the classification methods ranged from 0.660 for twelve-type LLA classification to 0.882 when classifying the three activity classes for the bigger SCI group. By dividing the twelve activity types into three classes, the performance of both the activity classification and EE estimation parts showed improved performance. Despite tree-based methods tendency to do so, the activity type prediction models showed no particular bias towards the majority classes when analysing the precision scores.

Two variants of activity types were used. First by using the twelve activity types that were recorded during the data collection, and then by dividing the ADLs into two classes, based on propulsion or no propulsion, which resulted in better performance of the activity type classification. The advantages for the overall EE estimation were however very small (see Table 12 and 13) and mostly statistically insignificant. Two ways of training the individual specialised regression models were also tested. They could either be trained on the actual activities or on the activities that were predicted by the classification method. Using the actual activity showed slightly better results, although only significant in the three-class version of the model in the SCI group.

When using the LLA group, the performance of the activity classification was considerably worse. The differences in results of the EE estimation methods were all deemed insignificant due to a small number of observations and small differences in MAPE scores.

All in all, the activity-specific model that was introduced in this thesis performed marginally better than the baseline regression model for the SCI group. The differences in performance were small and more optimisation and testing is needed to be able to consider the use of activity classification to estimate EE an improvement.

5 Discussion

As it is now, this activity-specific method brings no practical improvements over a more basic EE estimation method, like our baseline model. We have shown that improvements are definitely possible and that by using three activity classes the activity-model performed significantly better in the SCI group, with a MAPE drop from 0.367 to 0.350. Most versions of the activity-specific method showed minimal performance differences from the baseline model or no significant difference at all.

The classification model was substantially more accurate when classifying the three classes instead of the twelve activity types, but that brought little to no performance increase in the EE estimation. This could be because of a combination of two things. The first possibility is that the mistakes the classification algorithm makes are between activities that are already similar in movement patterns and EE, so it does not really matter as much when the activity classification is not very accurate. The second possibility is that the activities in the same class do not always have matching propulsion and EE patterns, which would mean the 'specialised' regression models are not really specialised in this case.

The data from the DACT-Wheel project was gathered in a controlled environment, which may not represent movement patterns for free living situations. People are often doing multiple things simultaneously, like watching a video, talking with someone and doing the dishes. This does not match the controlled lab environment where the subjects only did a small set of specific activities, with no more than one activity at a time and with a resting period inbetween. Further studies could train activity-specific models on data that is collected in a way that is more true to real life to see how they would perform.

As was stated in the introduction, the goal is to eventually develop a model that can estimate EE in real-time for wrist-worn wearables. With the way we implemented our model in Python, this is not possible, although there is nothing that would keep someone from developing a similar model to the one described in this thesis, but for real-time use. Every time EE is predicted, features can be calculated from the previous samples.

There are two main ways to increase the statistical significance of the performance of the activity-specific method compared to that of the baseline regression method. Assuming that there is an actual difference in performance, a bigger test set with more subjects would result in more degrees of freedom and thus a higher confidence. The second way would be to improve the performance difference by improving the performance of the activity-based method by studying potential improvements.

5.1 Future research

Since the scikit-learn's default hyperparameters for the random forest models were deemed adequate, only basic tuning was done. The maximum depth of the trees in the random forests was set to 12 to reduce overfitting and the number of trees was set to 200 for a slight performance increase. Other hyperparameters that could be optimised to potentially improve performance are the maximum number of features per random subset of features, which increases variance but increases bias, and both minimum number of samples per split and minimum number of samples per leave to further reduce overfitting.

Another way to improve the performance of activity-specific EE estimation would be to try out other methods of classification and regression. Montoye et al. showed that artificial neural networks are very promising for energy expenditure estimation from accelerometer data, especially for wrist-worn accelerometers [MBHP17].

Different implementations of activity-specific EE estimation could also bring some improvements. One of the challenges of splitting the dataset into subsets based on activities is that the training set per regression model becomes significantly smaller, especially when classifying many (imbalanced) activities. A way to convert the categorical activity types into continuous values is by using a *one-hot encoder*. This creates a binary array for every activity with a 1 on the place of that activity in the array and all others as 0. In this way, there is no need to split the dataset into subsets, because this can be fed into a single regression model.

As was discussed in section 1.2, some form of MET lookup table has also shown potential for non-wheelchair users, but more research on MET in wheelchair users still needs to be done to form an extensive and reliable list of MET values per activity. Besides linear regression, MET lookup tables and activity-specific regression, Albinali et al. also experimented with a custom MET lookup method [AIHR10]. This method uses a linear model which is trained to output the EE based on age, height, weight and resting heart rate. Before the start of the DACT-Wheel project, each participant filled in a questionnaire and many measurements were done, so a lot of data is available to build such a linear model for future studies.

Another approach to improve the performance of both the activity type classification and of the specialised EE estimation models, is by researching what activity classes could results in the most optimal performance. Less activity classes can improve the accuracy of the classification part, put the EE estimation will be less specialised. On the other hand, more activity classes might make the regression models more specialised, but it will also lead to smaller subsets and less specialised regression models. The difference in performance we saw was in most cases insignificant.

References

[Act15] Activinsights. Geneactiv brochure. https://www.activinsights.com/wp-content/uploads/2015/11/GENEActiv-Brochure-2015.pdf, 2015.

[AHW$^+$00] Barbara E Ainsworth, William L Haskell, Melicia C Whitt, Melinda L Irwin, Ann M Swartz, Scott J Strath, WILLIAM L O Brien, David R Bassett, Kathryn H Schmitz, Patricia O Emplaincourt, et al. 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, 2000.

[AIHR10] Fahd Albinali, Stephen Intille, William Haskell, and Mary Rosenberger. Using wearable activity type detection to improve physical activity energy expenditure estimation. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 311–320, 2010.

[BP04] Andrea C Buchholz and Paul B Pencharz. Energy expenditure in chronic spinal cord injury. *Current Opinion in Clinical Nutrition & Metabolic Care*, 7(6):635–639, 2004.

[CB11] Scott A Conger and David R Bassett. A compendium of energy costs of physical activities for individuals who use manual wheelchairs. *Adapted Physical Activity Quarterly*, 28(4):310–325, 2011.

[CCMH15] Patrick Carrington, Kevin Chang, Helena Mentis, and Amy Hurst. " but, i don't take steps" examining the inaccessibility of fitness trackers for wheelchair athletes. In *Proceedings of the 17th international acm sigaccess conference on computers & accessibility*, pages 193–201, 2015.

[CKH$^+$10] Scott E Crouter, Erin Kuffel, Jere D Haas, Edward A Frongillo, and David R Bassett Jr. A refined 2-regression model for the actigraph accelerometer. *Medicine and science in sports and exercise*, 42(5):1029, 2010.

[EPS$^+$15] Carly S Eckard, Alison L Pruziner, Allison D Sanchez, et al. Metabolic and body composition changes in first year following traumatic amputation. *Journal of rehabilitation research and development*, 52(5):553, 2015.

[FG20] Gary J Farkas and David R Gater. Energy expenditure and nutrition in neurogenic obesity following spinal cord injury. *Journal of physical medicine and rehabilitation (Wilmington, Del.)*, 2(1):11, 2020.

[FSGJ21] Gary J Farkas, Alicia Sneij, and David R Gater Jr. Energy expenditure following spinal cord injury: A delicate balance. *Topics in Spinal Cord Injury Rehabilitation*, 27(1):92–99, 2021.

[GKC$^+$05] E Garshick, A Kelley, SA Cohen, A Garrison, CG Tun, D Gagnon, and R Brown. A prospective assessment of mortality in chronic spinal cord injury. *Spinal cord*, 43(7):408–416, 2005.

[HYH$^+$ew] D Hoevenaars, I Yocarini, J F M Holla, S de Groot, W Kraaij, and T W J Janssen. Accuracy of heart rate measurement by the fitbit charge 2 during wheelchair activities in people with spinal cord injury: Instrument validation study. (under review).

[LCCC10] Wei Liu, Sanjay Chawla, David A Cieslak, and Nitesh V Chawla. A robust decision tree algorithm for imbalanced data sets. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 766–777. SIAM, 2010.

[MBHP17] Alexander HK Montoye, Munni Begum, Zachary Henning, and Karin A Pfeiffer. Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. *Physiological measurement*, 38(2):343, 2017.

[MLK07] Jonathan Myers, Matthew Lee, and Jenny Kiratli. Cardiovascular disease in spinal cord injury: an overview of prevalence, risk, evaluation, and management. *American journal of physical medicine & rehabilitation*, 86(2):142–152, 2007.

[NRTB17] Tom E Nightingale, Peter C Rouse, Dylan Thompson, and James LJ Bilzon. Measurement of physical activity and energy expenditure in wheelchair users: methods, considerations and future directions. *Sports medicine-open*, 3(1):1–16, 2017.

[NWTB14] Tom Edward Nightingale, Jean-Philippe Walhim, Dylan Thompson, and JAMES L Bilzon. Predicting physical activity energy expenditure in manual wheelchair users. *Med Sci Sports Exerc*, 46(9):1849–1858, 2014.

[PBCC96] AM Prentice, AE Black, WA Coward, and TJ Cole. Energy expenditure in overweight and obese adults in affluent societies: an analysis of 319 doubly-labelled water measurements. *European journal of clinical nutrition*, 50(2):93–97, 1996.

[PGCB16] Toby G Pavey, Sjaan R Gomersall, Bronwyn K Clark, and Wendy J Brown. The validity of the geneactiv wrist-worn accelerometer for measuring adult sedentary time in free living. *Journal of science and medicine in sport*, 19(5):395–399, 2016.

[PRB$^+$18] Werner L Popp, Lea Richner, Michael Brogioli, Britta Wilms, Christina M Spengler, Armin EP Curt, Michelle L Starkey, and Roger Gassert. Estimation of energy expenditure in wheelchair-bound spinal cord injured individuals using inertial measurement units. *Frontiers in neurology*, 9:478, 2018.

[She03] Roy J Shephard. Limits to the measurement of habitual physical activity by questionnaires. *British journal of sports medicine*, 37(3):197–206, 2003.

[SLJ$^+$17] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, and Konrad P Kording. The need to approximate the use-case in clinical machine learning. *Gigascience*, 6(5):gix019, 2017.

[SMGT13] Shane N Sweet, Kathleen A Martin Ginis, and Jennifer R Tomasone. Investigating intermediary variables in the physical activity and quality of life relationship in persons with spinal cord injury. *Health psychology*, 32(8):877, 2013.

[SVHD20] Yousif J Shwetar, Akhila L Veerubhotla, Zijian Huang, and Dan Ding. Comparative validity of energy expenditure prediction algorithms using wearable devices for people with spinal cord injury. *Spinal cord*, 58(7):821–830, 2020.

[VHGDL+13] Vincent T Van Hees, Lukas Gorzelniak, Emmanuel Carlos Dean León, Martin Eder, Marcelo Pias, Salman Taherian, Ulf Ekelund, Frida Renström, Paul W Franks, Alexander Horsch, et al. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PloS one*, 8(4):e61691, 2013.

[Wei49] JB de V Weir. New methods for calculating metabolic rate with special reference to protein metabolism. *The Journal of physiology*, 109(1-2):1–9, 1949.

[WFH11] H. Ian Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques.* Elsevier, third edition, 2011.