Opleiding Informatica & Economie

Classifying strength exercises performed by wheelchair users using accelerometer data of wearables

Lorenzo Spierings

Supervisors:
Dr. I.E. Yocarini & Dr. M. van Leeuwen

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl 15/07/2020

Abstract

Inadequate physical activity increases the risk of various health problems. For wheelchair users, physical activity can be more challenging and daily goals are often not met. Wearables and e-platforms do not offer wheelchair users the same benefits as non-wheelchair users. With the help of accelerometer data obtained by multiple accelerometer devices, we studied human activity recognition (HAR) for strength exercises performed by 40 wheelchair users with a spinal cord injury (SCI) or lower limb amputation (LLA). The classification of seven strength exercise categories was done with supervised machine learning techniques including a recurrent neural network (RNN). To determine the performance of the classifier, we compared the performance between classifying strength exercises with daily activities. In addition, we tested our models to injury-specific subgroups to detect diversity between different types of wheelchair users. With these experiments, we concluded that creating a classification model for strength exercises in a free moving fitness environment is challenging. Wheelchair users tend to perform exercises with different execution forms, due to variation in physical limitations. Our model is not accurate enough for a reasonable adjustment to energy expenditure estimation with an accuracy score of 0.55. Improvement of the model within the field of strength exercises among wheelchair users is necessary before it can actually be deployed to wearables and e-platforms.

Acknowledgements

This thesis is written as part of the bachelor degree Computer Science & Economics at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University. I would like to give a special thanks to my supervisors, Iris Yocarini and Matthijs van Leeuwen, for their intensive support and valuable feedback during this thesis project.

Contents

1 Introduction

In order to pursue a healthy lifestyle, it is important to get enough exercise every day as it is correlated with general health. Unfortunately, for wheelchair users physical activity may be more challenging and daily targets are often not met, leaving them at risk for various physical and mental health problems such as obesity, cardiovascular diseases, hypertension, diabetes and some forms of cancer. [21, 14, 32] One reason being, the fact that performing daily activities is more difficult and energy-intensive compared to non-wheelchair users. [33, 42] As a result, wheelchair users are generally more tired, leaving even less energy to achieve a healthy amount of daily physical activity, causing them to enter a vicious circle, in which physical activity goes down while body weight goes up. Especially for wheelchair users, physical activity is important though, given their higher risk of health problems. [40, 32] The goal of this thesis is to take a step in the direction of a healthier lifestyle of wheelchair users. [13] In the pursuit of a healthy lifestyle, tools in the form of wearables and e-platforms may be helpful.

Wearable fitness devices and e-platforms aim to improve general health by stimulating physical activity. [5] Assistance is then given to the user by tracking their activity and reporting the progress towards their physical activity goal, resulting in physical and mental health improvements. [5]. E-platforms and wearables provide recommendations for a sufficient amount of daily activity, based on the total energy expenditure (TEE) estimation. TEE is the daily amount of calories burned by the human body adjusted to the amount of activity. [33] Physical activity is the most determining component of the TEE, explaining why health goals are often measured by the amount of physical activity. It has been shown that a TEE estimation is more accurate when an activity classification is included in the estimation process, specifying a different estimation function per type of activity. For activity classification, human activity recognition (HAR) is used to identify a specific type of physical activity performed, by using a series of accelerometer data of the movement. [24] Unfortunately, wheelchair users do not benefit as much from these e-platforms and wearables compared to non-wheelchair users, due to their restriction of movement. Furthermore, a movement performed in a wheelchair involves different sensor patterns, resulting in reduced accuracy of the built-in basic activity classification of most wearables. [7] In addition, the pedometer is inaccurate because no actual steps are taken. [7] These limitations will eventually lead to inaccurate advice regarding the sufficient amount of physical activity. The end goal of this study is to increase the accuracy and usability of these wearables and e-platforms amongst wheelchair users. For this purpose a HAR model is build for wheelchair users.

To strive for a better energy expenditure estimation for wheelchair users, a reliable performing classification model is required. [35] Strength training is proven to benefit general health, prevent injuries and also makes performing daily activities easier for wheelchair users. [2] HAR based on accelerometer data has already been studied for various healthy populations such as younger adults [20] and the elderly. [30] Besides healthy populations, HAR for wheelchair users with an incomplete spinal cord injury has also been studied. [35] However, the aforementioned studies focused on recognizing daily activities, without including strength exercises. This study will exclusively focus on HAR for strength exercises performed by wheelchair users, since there are no previous HAR studies including wheelchair users focusing on strength activities. By using Data Science to create this classifier we hope to indirectly contribute to an increase in accuracy of energy expenditure estimation and consequently the usability of wearables among wheelchair users performing strength training.

1.1 Research Question & Challenges

The data used for the classification model was collected in a non-controlled, free moving fitness setting as a part of the DACT-Wheel research project. This project was targeting wheelchair users suffering from either a spinal cord injury (SCI) or lower limb amputation (LLA). During this collection process, strength exercises as well as daily activities were performed. [19] This allows us to compare the daily activity classification model with results from previous studies. Nevertheless, the main focus will be on the strength data, i.e., to create a classification model of strength exercises.

The participants were free to choose the strength exercises, which leads to a challenge in analyzing and processing the data. The participant and the researcher were the persons documenting the exercise name. Hereby, each participant can interpret the name of the exercise differently, while being the same movement. Otherwise, the researcher could name an exercise in an alternative way, which includes their interpretation. The fact that not all wheelchair users are limited in their movement to the same extent, results in a great variety in the form of the exercises. Another challenge in building the classifier is that the same signals may be linked to multiple exercises, as different exercises may contain the same direction of movements.

The study aims to contribute to better usability and accuracy of fitness wearables and e-platforms, with the help of HAR of strength exercises for wheelchair users. Therefore, the central research question is represented as:

"How can we build a classifier of strength exercises performed by wheelchair users based on accelerometer and heart rate data."

To support the main research question, several sub-questions will be examined:

- To which extent does a neural network perform better in classifying strength exercises than more traditional machine learning techniques?
- What are the differences between predicting strength exercises and daily activities?
- What are the performance differences when looking at the different types of wheelchair users?

1.2 Thesis outline

In section 2 we will start with explaining background information. Followed is section 3, where the data will be discussed. Section 4 will be used to describe the challenges encountered in the preprocessing step as well as the methodology for the experiments. The results of these experiments will be discussed in section 5. The final section of this thesis, section 6, contains the conclusions and a discussion about future research opportunities.

2 Background

This section will discuss the background of the DACT-Wheel project. The impact of spinal cord injury (SCI) and lower limb amputation (LLA) on physical activity will be reviewed. Also, a general overview of human activity recognition (HAR) will be discussed.

2.1 DACT-Wheel

DACT-Wheel is a research project with a duration of four years, which started in 2018. The main goal of the DACT-Wheel project is to help wheelchair users in achieving a healthy lifestyle. For the general population, wearables like the Fitbit and e-platforms like Virtuagym are used to contribute to a healthy lifestyle. Virtuagym provides applications for nutritional and workout guidance, which help people to obtain a sufficient amount of physical activity. The nutritional guidance tool needs adjustment to a wheelchair user's energy expenditure estimation, to accurately help this specific user group. [43]

With the help of various devices, a large amount of sensor data was collected. This data consists of tri-axial sensor data during the performance of strength exercises and daily activities. The participants were wheelchair users with a spinal cord injury (SCI) or a lower limb amputation (LLA). In addition to the aforementioned injuries, there was also a small group of non-wheelchair users in a wheelchair. More detailed information about the type of data and collection process will be discussed in section 3.

1. **Spinal Cord Injury (SCI)**
   A spinal cord injury means that any part of the spinal cord is damaged. The impact of this injury can affect many areas of a person's life. The loss of control of your limbs is the most drastic consequence of the SCI injury. [27] Also, symptoms of spasticity are likely to occur. [1] These are only two of the many effects of SCI. Overall the limitation of movement is the biggest culprit when it comes to physical activity. The combination of the aforementioned effects is the reason that most SCI patients generally end up in a wheelchair.

2. **Lower Limb Amputation (LLA)**
   A lower limb amputation comes in different extremity levels. However, all kinds of amputations do have an impact on a person's life quality. Many people suffer from negative psychological effects like depression, as well as physical pains. [17] The adjustment to limb loss is different from person to person, but often results in a life in a wheelchair. The impact on the amount of physical activity (PA) after LLA is drastic, whereas PA requires additional strength and energy. Furthermore, the flexibility in exercise choice is limited, due to the need for specialized equipment like prosthesis. [23]

A goal of the DACT-Wheel project is to create a wheelchair-adapted energy expenditure estimation model. Energy expenditure (EE) is considered the amount of energy a person uses on a daily basis. The total energy expenditure (TEE) is consisting of three main components. [33, 16] The first and biggest component is the resting energy expenditure (REE), which indicates the basic expenditure to keep the vital functions of the body running. The second component is diet-induced thermogenesis (DIT), determined by a person's nutritional intake. The last component is physical activity (PA), controlled by the amount of movement leading to a thermic effect on the energy expenditure.

Wheelchair users with a body that is partly paralyzed or has amputated limbs have reduced control of temperature regulation and vascular systems. The reduced control leads to an alternative heart rate and blood pressure pattern, which generally increases much faster than the general population. [18] The unconventional body regulation of wheelchair users results in an alternate REE component. To reduce the risks of diseases, especially for wheelchair users, the PA component must therefore be as high as possible.

Physical activity in the form of exercise is proven to be beneficial for general health, especially for wheelchair users, who already have a higher risk of diseases. [14, 40] However, multiple factors lead to more challenges to get enough PA in for wheelchair users. Wheelchair users are also more prone to injuries and other health issues. The load on the upper body during daily tasks is significantly higher than for non-wheelchair users, therefore exercising is shown to be important in preventing injuries. [31] With PA being the most variable component of the energy expenditure, a classification model of strength exercises performed by wheelchair users is shown to be helpful towards a more accurate estimation of the energy expenditure. [35] The classification models for these purposes are known as human activity recognition (HAR).

2.2 Human Activity Recognition

Human activity recognition (HAR) is known as a computational process of recognizing an activity from a sequence of data. The goal is to get knowledge out of data obtained from sensors. [22] Wearables like smartwatches are equipped with tri-axial accelerometer and heart rate sensors to obtain data. HAR is mainly based on accelerometer data, which is often supplemented with other types of data. [45, 22] Physiological inputs, such as heart rate, are often considered to be included in the HAR process but are not always helpful for the classifier. Tapia et al. showed that the heart rate remains relatively high after performing an intense activity and is therefore not always representative of the activity. [39] This factor can be eliminated by taking into account the longitudinal aspect of the data, for example by adjusting window size in sequences or including a lag in the model. Finally, there are environmental attributes that are regularly included in HAR, such as temperature or humidity. However, these environmental features are only used in combination with accelerometer data when creating a classifier. [26]

After the data collection part, the data needs to be processed to build a HAR model. During preprocessing, the sampling frequencies are matched and features are extracted. When the data is preprocessed, the recognition model can be build by learning from the features using supervised machine learning techniques including neural networks. After this learning phase, the model can then predict new instances and classify the activity performed. [20, 24, 26, 22]

Popp et al. studied HAR for daily activities, including wheelchair users. [35, 34] Their classification model is accurate, by using a traditional machine learning technique. When using a neural network, results showed that the run-time length often correlates with the accuracy of the model. [4]. Also, the quality of the data and the number of classes influence the accuracy of a model.

3 Data

During the DACT-Wheel project, data was collected in a non-controlled, free moving fitness setting. Here, a sample group performed daily activities and strength exercises. The participants were monitored by various wearable measuring devices during these activities. In this section, we will discuss the data.

3.1 Samples

For the data collection, a sample group has been selected based on criteria. The group consists of spinal cord injury (SCI) individuals, lower limb amputation (LLA) individuals, and non-wheelchair users in a wheelchair. To participate in the study, the participants must have an age between 18 and 75. Besides the age criterion, the participant must be dependent on a manual wheelchair and the injury (SCI/LLA) must be chronic ($>$ 1 year). [19] The final sample group consists of 62 individuals, whereby the largest group is represented by wheelchair users with a spinal cord injury. The exact distribution of the original sample group is specified in the second column of Table 1.

Not all measurements on the sample group resulted in a complete data set. For this reason, there are persons whose data has not been included in the model. The main reasons for not including a person's data are either missing complete data or data of poor quality. For example, it may have happened that a device has failed to record the measurement. In one specific case, the device was mounted on the wrong wrist, which also can negatively affect the model. In addition to technical failure, some participants were unable to resume exercising, due to injuries. The third column of Table 1 shows the final number of participants per group, whose data was included in our model.

Participant Type	**Tested amount**	**Final amount**
SCI	39	25
LLA	13	10
Non-wheelchair user	10	5

Table 1: Sample group distribution before and after removing poor quality data.

3.2 Activities

The activities for which data is collected can be divided into three different categories. The first measurement was during a resting state, which we do not use in our research. The second measurement phase was during the performance of at least eleven daily activities (ADL). [19] These activities are described in more detail in Table 3. This data will be used in our experiment to compare the performance of classifying daily activities (ADL) compared to strength exercises. In this case, exclusively the data where daily activities were performed, will be used.

In the third phase of the collection process, the data of performing strength exercises was obtained. This data is the most interesting data for this particular study. Each participant performed for 30 minutes strength exercises, in which three to five strength exercises of own their own choice were done. The different exercises that were performed and their popularity amongst the participants are shown in Table 2. Each exercise selected was performed in three sets. The number of repetitions per set was chosen by the participant, and involved in most cases twelve repetitions. Also, the exercise weight was chosen by the participant, which could be either a set of dumbbells or a theraband. The dumbbells varied in weight from two to fourteen kilograms. The therabands were available in two different versions, heavy and medium, which could also be combined during an exercise. Besides using therabands or dumbbells, exercises could also be performed with bodyweight only. The participants were free to move in between the strength exercises, resulting in a collection of random movements. Therefore, the data in between the exercises can be considered as noise and was omitted.

Exercise Type	**Count**
Biceps Curl	50
External Rotation	32
Reverse Fly	27
Internal Rotation	27
Front Raise	24
Triceps Extension	20
Shoulder Press	19
Side Raise	9
Dumbbell Row	9
Seated Row	8
Triceps Kickback	7
Dumbbell Press	4
Chest Press	4
Chest Fly	1

Table 2: Exercises performed by the sample group.

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 3: Daily activities performed by the sample group. [19]

3.3 Devices

For the data collection, multiple devices were used. According to a study by Benedetto et al., the measured values of wearables often contain a measurement error and can deviate from the real values. The reason for the deviation has to do with different measuring methods of the devices for signals like heart rate. [3] Complementing to the different measuring methods, the measurable value is a non-observable value, which makes it even harder to decide which method is the best. When deviating values are used by the classifier, it may have difficulties with new predictions. Therefore, multiple devices were used for the same measurements, to train the classifier without relying on one device.

An overview of the used measuring devices with their main purpose is given in Figure 1.

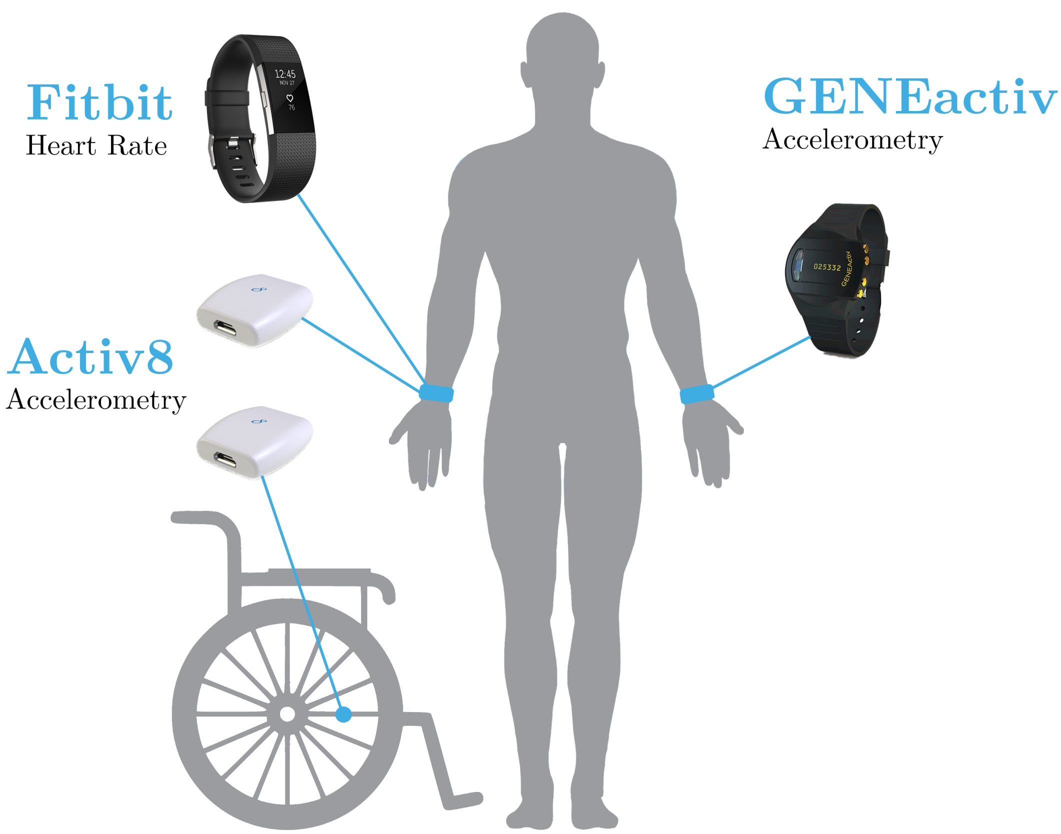

Figure 1: Overview of the measuring devices used and their locations.

1. **Activ8**
   The Activ8 is a commercially available accelerometer device, normally worn on the frontal thigh. This small device is used regularly in clinical environments, because of the low cost and user-friendly dashboard. The Activ8 is proven to be able to classify basic movements and energy expenditures for healthy people. [12] However we only use the raw accelerometer data captured with the Activ8 for this study.

   The Activ8 contains a tri-axial accelerometer, which publishes raw x, y and z values with a sampling rate of 12,5 Hz. [10] The manufacturer has calibrated the device for the frontal tight, to ensure the correct functioning of the earlier mentioned classification function of basic movements. This specification was intentionally ignored during the project because the raw accelerometer of specific limbs, such as the wrist, was of more importance in this study. Furthermore, strength exercising in a wheelchair involves a lot of arm movements.

   Two Activ8 devices were used in this study. One of the devices was worn around the right wrist. The second device was attached to the wheel of the wheelchair to register the movement of the wheel. This placement was chosen to register the highly fluctuating movement of the wrist and wheel of the wheelchair while performing a strength exercise.

2. **GENEactiv**
   The GENEactiv is the second accelerometer device used in this research, to be worn on the wrist. This device also contains a tri-axial meter similar to that of the Activ8. However, this device is capable to record raw accelerometer data at a frequency up to 100 Hz. [36] During the data collection, the GENEactiv was worn on the left wrist.

3. **Fitbit**
   The Fitbit is a consumer-grade smartwatch, focused on health purposes. In this research, the Fitbit is mainly used for tracking the heart rate. To track heart rate, Fitbit uses PurePulse light-emitting diodes, a technique also known as PPG. The diodes continuously measure the blood volume change on the skin surface. With these constant measurements of blood volume changes, the Fitbit ultimately estimates the heart rate. [3] Next to the heart rate measurement, the Fitbit is used for its physical activity count feature. [19]

4 Methodology

This section will be about the methods used in this research. The preprocessing methods as well as the experimental setup will be discussed.

4.1 Preprocessing

In order to create a supervised machine learning model, the data must be preprocessed first. This section covers the entire preprocessing process. Figure 2 shows a schematic of all steps before the model can be built.

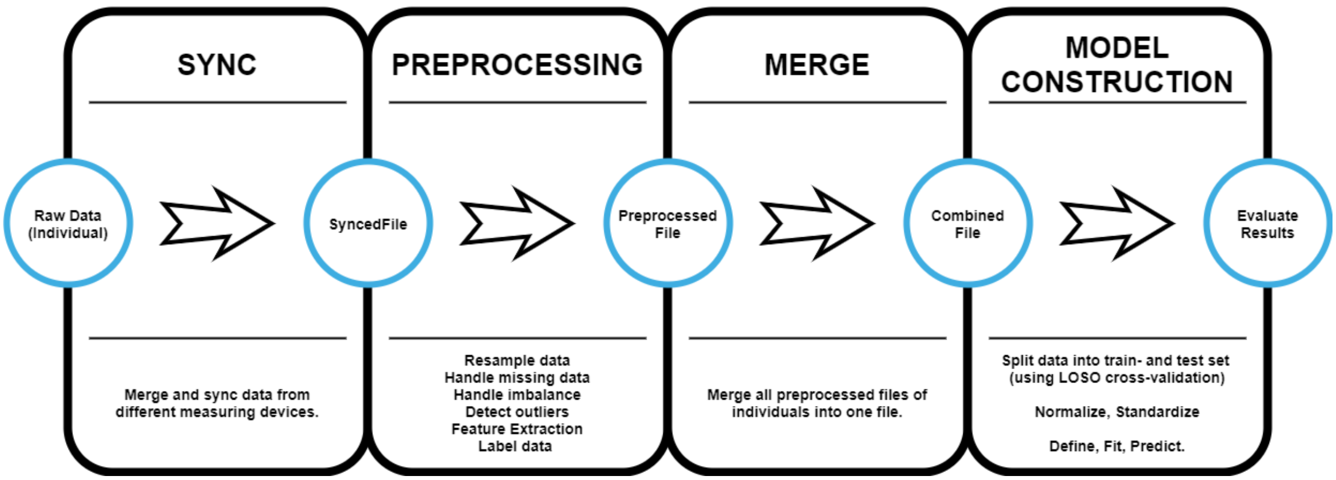

Figure 2: Overview of the complete process.

4.1.1 Synchronizing & Resampling

The first step is the synchronization of signals measured by the different devices. The accelerometer data from the GENEactiv and the Activ8 were not merged and needed to be aligned based on the corresponding timestamp. Because the Activ8 showed 32 measurements during a 2,5 seconds period, the sampling frequency is approximately 12,5 Hz. This frequency is confirming the frequency as specified by the manufacturer in section 3.3. We chose to match all other measurements to the sampling frequency of the Activ8. As a result, the GENEactiv has been downsampled from 100 Hz to 12,5 Hz. Since the Fitbit's sampling frequency compared to other devices, the data had to be upsampled. We used interpolation to upsample the data within known data points.

After the sampling rate of the two devices matched, we could merge the devices together. We tried two different approaches to merge the two devices. The first method being, merging based on the nearest timestamp. The nearest timestamp was determined based on the date-time index of both devices. With the merge function from the Pandas library, we managed to merge the two devices, by using the date-time index with the parameter merge on set to 'nearest'. In addition to the method of merging based on the nearest timestamp, we also considered merging the signals using the auto-correlation. The autocorrelation looks at a delayed version of itself, focusing on finding the greatest correlation of the signal's peaks. Ultimately, synchronization will then take place based on these matching peaks. [20, 6] However, the method based on the nearest timestamp seemed sufficient after the alignment was assessed by reviewing plots as shown in Figure 3. These plots were also used to detect missing data or a malfunctioning device.

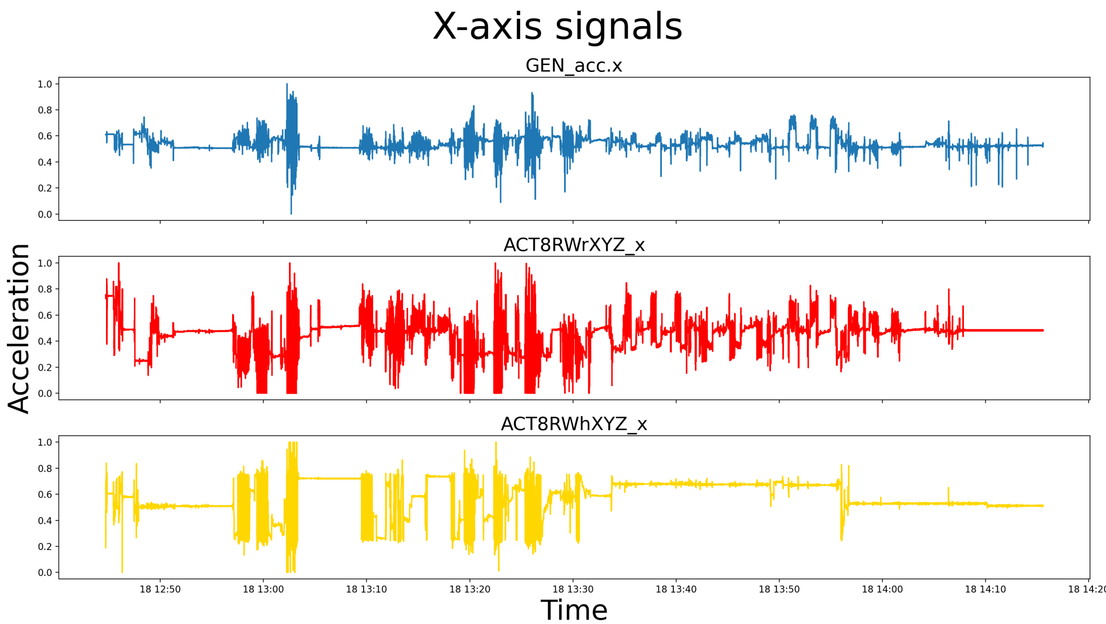

Figure 3: Plots of the X-axis measurements for person AA003 after syncing the devices.

4.1.2 Labeling

For this study, the data had to be labeled. The labels were added based on the timestamps. Timestamps were recorded by the researcher who took the measurement at the beginning and end of each activity. Labels to distinguish the type of activity (rest, daily activity, strength activity) were added initially. Since the focus of this research is mainly on strength exercises, the distinction between different strength exercises is important. The naming of the exercises were all recorded by the researcher, so interpretation was playing a role in naming the exercise. Therefore, a sanitized overview was created in which all the exercise names were normalized. Identical exercises were named differently, e.g., 'endorotation' and 'internal rotation'. After normalizing the naming of the exercises, we also recategorized the exercises due to the wide variety. Some exercises were performed only a few times, resulting in class imbalance followed by a higher bias of the classifier towards the majority class. An example is combining 'internal rotation' and 'external rotation' into one category, as the movement pattern is the same, with a slightly altered angle. This resulted in a reduction in the number of different exercise categories. In the end, only seven different strength exercises were left, based on the direction of movement and category of the exercise. The new recategorized strength exercise classes with the corresponding number of occurrences are shown in Figure 4.

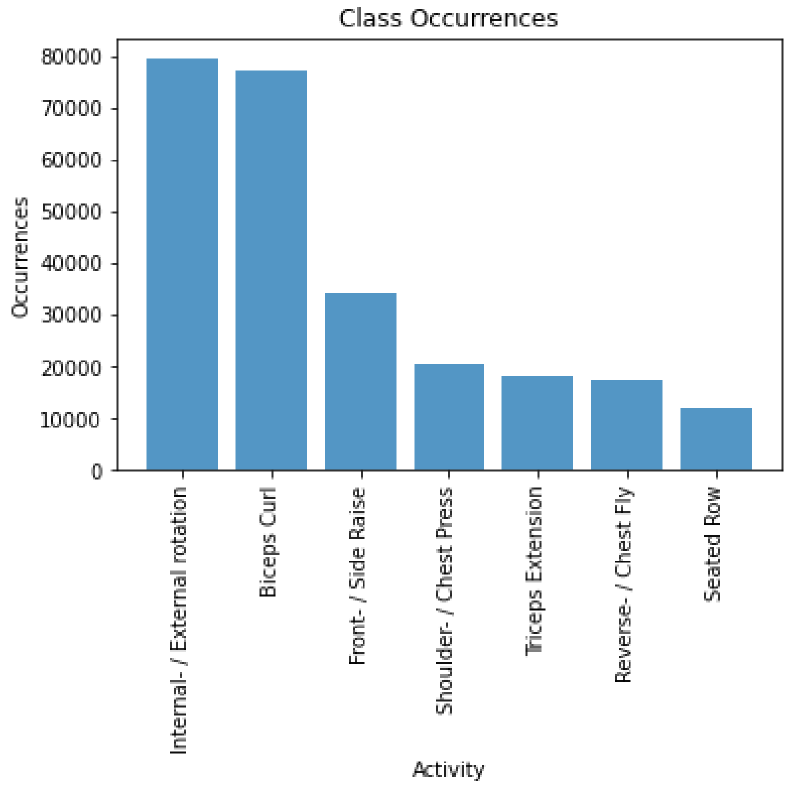

Figure 4: Class distribution after regrouping the exercises.

4.1.3 Feature Extraction

The measuring devices provide a large number of raw features, as mentioned earlier in section 3.3. The data of the classification of basic movements from the Activ8 is the result of a classifier from the company, which already is processed data. Non-raw features like the basic classification movements of the Activ8 were left out for two reasons. First of all, we tried to select data as raw as possible. Secondly, to ensure optimal operation of the Activ8 classifier, the device should be mounted on the place where it was calibrated for, which was not the case for this study.

- **X, Y and Z acceleration** (normalized) from the GENactiv.
- **X, Y and Z acceleration** (normalized) from the Activ8 (Wrist-mounted).
- **X, Y and Z acceleration** (normalized) from the Activ8 (Wheel-mounted).
- **Heart rate** (normalized) from the Fitbit.
- **Steps / Distance** from the Fitbit.
- **Lux** (amount of light) from the GENactiv.
- **Temperature** from the GENactiv.

Several features were constructed as a derivation of the raw measurements. Aggregation of statistical properties like *avg, med, max, min, std* captures a lot of properties of the data within a time window. These features were extracted after reviewing research conducted by Cachucho et al., about multivariate time series with mixed sampling rates. [6] Besides aggregation features, Khan et al. used magnitude derived from the raw X, Y, Z acceleration. [20]. The complete listing of added features is available in Table 9 in Appendix A.

- **avg, med, max, min, std** for X, Y and Z-axis of GENactiv. The value was calculated within a 8s time window.
- **Signal Vector Magnitude** of the GENactiv and both Activ8 devices. The magnitude was calculated using the following formula: $\sqrt{(x^2 + y^2 + z^2)}$
- **Velocity** of the GENactiv and both Activ8 devices. Velocity is derived from Magnitude, and has the advantage that the direction of the signal remains intact.

Figure 16 in Appendix B shows the correlation between the included features. The matrix clearly shows that the features with the highest correlation are derived from the corresponding axis signal. The rest of the features are generally less correlated to each other. In the end, the final complete data set contains 236043 samples, whereby each sample consists out of 32 features.

4.1.4 Neural Network Specific Steps

Extra steps are needed when it comes to preprocessing the data for the neural network. Recurrent neural networks can handle sequential data, and therefore we created sequences of a specified length. Also, normalization of the whole data set is beneficial for a neural network, instead of only applying it to accelerometer and heart rate data for the baseline models. The reason for scaling the whole data set is that most activation functions of ANN produce an output between [0,1] or [-1,1]. This way the learning process is not suffering from incomparable values. [37]

4.2 Model Construction

Machine learning is used to predict the type of strength exercise performed by a wheelchair-user, based on sensor data. In this research, we use supervised machine learning techniques, which means that the data contains labels. This way the model can learn how the signals of the person's arms as well as the wheelchair acceleration correspond to the performance of a specific strength exercise. After the learning phase, the model can be applied to classify the movement of the test set, containing one person's unlabeled data.

To perform machine learning, the preprocessed data have been split by person. After splitting the data, Leave-One-Subject-Out (LOSO) cross-validation has been applied on 100% of the data set. This means that for `n` persons, `n` folds are applied. Each fold one person is forming the test set, and `n-1` persons are forming the training set. By calculating the average over all folds, we can evaluate the performance of the model. The choice to use LOSO cross-validation is to eliminate an accidental preference of participants in the test group for which the model performs better than the average participant. Some participants may be more familiar with an exercise form than other participants, resulting in a better recognizable movement pattern and corresponding accuracy. Another advantage of LOSO is that the data of different people remains separate and is not mixed in the train and test set. Otherwise, data samples from the same participant can occur in both the training and test set and data linked to a person provides valuable information that otherwise would be lost when mixing the data. [28] LOSO ensures that each participant will be the test person once, thus the average of all accuracies determines the overall performance level of the model. The individual scores can also be used to detect outliers amongst the participants. Earlier research conducted by Okai et al. and Popp et al. on HAR with sensor data, used a similar LOSO approach. [25, 30, 35]

```
         Example of LOSO with n = 7
Fold 1: TRAIN: [1 2 3 4 5 6] TEST: [0]
Fold 2: TRAIN: [0 2 3 4 5 6] TEST: [1]
Fold 3: TRAIN: [0 1 3 4 5 6] TEST: [2]
Fold 4: TRAIN: [0 1 2 4 5 6] TEST: [3]
Fold 5: TRAIN: [0 1 2 3 5 6] TEST: [4]
Fold 6: TRAIN: [0 1 2 3 4 6] TEST: [5]
Fold 7: TRAIN: [0 1 2 3 4 5] TEST: [6]

Accuracy = (score_1 + score_2 + score_3 + score_4 + score_5 + score_6 + score_7) / 7
```

Performance metrics will be discussed in more detail in section 4.4.

4.2.1 Baseline Models

Whereas Kahn et al. researched human activity recognition (HAR) containing a likewise data signal from an accelerometer, they were predicting HAR for non-wheelchair users. In this research, several classification algorithms were compared to each other. [20] Popp et al. also did HAR research for iSCI participants, using the KNN classifier. [35, 34] Based on these studies, the following baseline classification methods were selected for this study (using default parameters):

- **Decision Tree Classifier**
  The decision tree classifier has a flow chart structure of nodes and can be seen as a tree-like model of decisions. A decision tree exists out of decision nodes and leaf nodes, where the starting node is (root node) is a decision node. Each node is connected via branches to another node, becoming more informative when a new layer of the tree is reached. Data will be split on a feature at each decision node, leading to multiple branches. To choose the feature to split on, the algorithm looks for the feature with the highest information gain. Information gain is the decrease in entropy after a split, which can be seen as a measure of uncertainty. [38] At the end of each branch reaches a leaf node is reached, representing a class label for the given data point. [8]

- **K-Nearest Neighbor Classifier (KNN)**
  The KNN classifier determines the class of a single data point by looking at k data points around it. The total number of neighbors to look at is defined by k and needs to be chosen in prior. The distance to each surrounding data point is calculated with the help of a distance function such as the euclidean distance. In the end, the current data point gets the same class label as the surrounding data points with the smallest distance. [41]

- **Random Forest Classifier**
  The random forest classifier makes use of multiple decision tree classifier instances. The outcome is an aggregation of multiple trees, making the classifier less prone to overfitting. In addition to being more resistant to overfitting, the use of multiple decision trees makes the model more accurate when parts of the data are missing. The data will be split into various sub-samples of the total data set. Each tree in the random forest predicts a class, after which the most predicted class will be the model's prediction. [44]

4.2.2 Neural Network

In comparison to regular machine learning approaches using baseline models, Kahn et al. showed that artificial neural networks (ANN) are more accurate in HAR including only non-wheelchair users. [30] Okai et al., showed a comparison between two types of ANNs, targeting the population of the elderly. Their study showed that recurrent neural networks (RNN) with the GRU type layers are the best performing models for HAR. [30]. Popp et al. also used ANN's for HAR of daily activities on incomplete SCI participants. However, their findings were that an ANN performed less well for their use case than a regular machine learning classifier. They assumed that the ANN might outperform the base models if more participants are included. Their reason being that the current population was very heterogeneous with relatively fewer outliers. while assuming that ANN's weigh outliers heavier. [35].

- **Recurrent Neural Network (RNN)**
  An artificial neural network (ANN) consists of layers of neurons, which can be seen as the processing units of the network. The network starts with an input layer and ends with an output layer. In between these input and output layers, multiple hidden layers are doing the computational work. All layers are connected with each other through channels with a corresponding weight. Each neuron has an activation function, which determines whether the neuron will be activated and pass the data through the next layer of the network or not. This pass-through of data through the network is called forward propagation. The input value of the activation function is a numerical value called the bias together with the sum of input values multiplied the corresponding channel weights. When the output layer is reached, the neuron with the highest value (probability) is the class predicted by the neural network. Then the network will compare the outcome to the actual output and calculate the error. The weights are then adjusted by feeding the information back to the network, also called backpropagation. This entire process is iteratively repeated until the network can predict the classes correctly nearly at all times. [15] A schematic illustration is visible in Figure 5.

  A recurrent neural network (RNN) is an ANN with a special type of neurons in the hidden layers. These neurons are like memory cells, with the special feature of handling sequential data more efficiently. These types of layers neurons are typically called Long-Term-Short-Memory Cells (LSTM), as well as Gated Recurrent Units (GRU). [9]

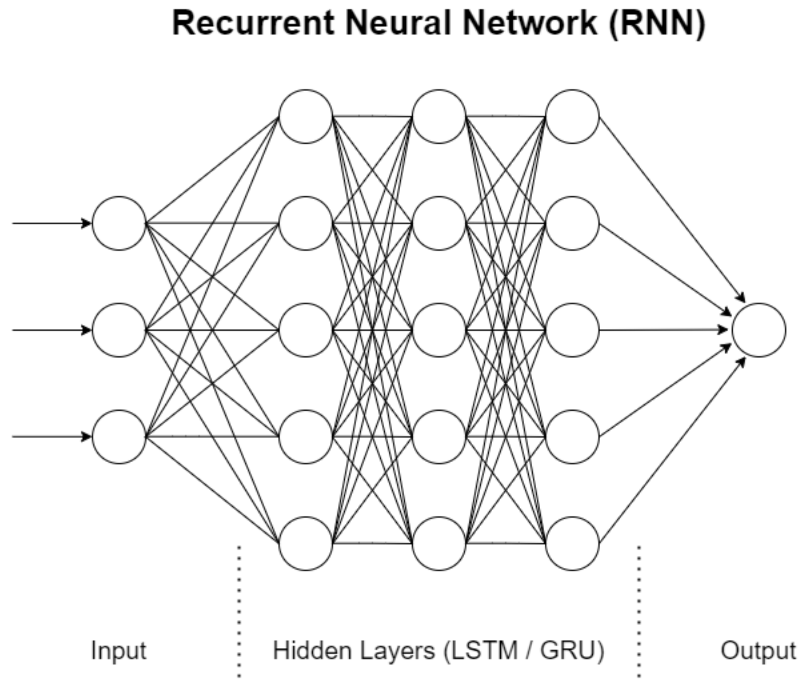

Figure 5: Schematic overview of a Recurrent Neural Network.

Based on the characteristic of handling sequences of data and previous successes of the RNN in HAR, we chose to use a RNN for our classification task. RNNs have relatively high computational costs and no default structure compared to traditional machine learning models. Therefore, we chose to try out and tweak the RNN structure in a more traditional setup, with a train, validation and test set. The validation and test set existed out of one randomly chosen participant from each group, which ensures that the model is not tuned to a specific type of wheelchair user. The validation set was used to check the actual performance while fitting the model and tweaking parameters, such as *EPOCHS, BATCH_SIZE, LAYER_NODES* and *SEQ_LEN*. The test set was created so that not all data is used during the tuning process. Furthermore, there was no actual testing done on this set, since the final test run was done using LOSO cross-validation.

EPOCHS indicates how many times the model is learning from the entire training data. *BATCH_SIZE* tells the model when to readjusted the weights in the network. *LAYER_NODES* are the number of nodes per hidden layer in the network. *SEQ_LEN* clarifies the number of data points included in one sequence, as RNN's expect sequences as input. After several parameter combinations had been tested, a trade-off between run-time and marginal improvements in accuracy was made. In the end, we made one final run using the LOSO cross-validation method, to be able to compare the performance to the traditional machine learning models.

4.3 Test Setup

To determine the differences in predicting strength exercise across different types of wheelchair users, we created four different data sets to run the model on. The first data set contains the complete data, including all types of wheelchair users. This complete data set will be used for the main experiment of classifying strength exercises, since the model has the most number of samples to train on. In addition to the strength exercises, a second experiment will be done where the models will be tested for classifying daily activities (ADL). This way we can illustrate whether the exercise form plays a big role. Daily activities tend to suffer less from diversity in the form than strength exercises while performing the activity.

The second and the third group containing respectively SCI and LLA wheelchair users only. The population of non-wheelchair users in wheelchairs is forming the last data set. A comparison between these groups defines the last experiment, providing an insight into the predictability of certain types of wheelchair users. Looking at the different injuries, participants with a certain injury are more likely to be restricted to the same extent as other participants with the same injury. The expectation is that the type-specific data sets may have higher accuracy compared to the complete data set due to less variation in different types of wheelchair users (SCI/LLA mixed). The fact that the type-specific data sets contain less data, means fewer samples for the model to train on. Fewer samples normally affects the model in a negative way in terms of accuracy. However, the expectation is that if the sample size is not extremely small (< 10 participants), the benefits of a population with the same type of injury will be greater.

The following version of tools are used to do the experiments: Python 3.8.1, Jupyter Notebook 6.0.3, Sci-kit Learn 0.22.1, Keras 2.2.4 and Tensorflow 2.1.0

4.4 Performance Metrics

To determine the performance of a model, we look at the accuracy of each model. We also plot the complete confusion matrices, to see which classes generally perform well and detect possible biases. The accuracy is calculated using the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FP). [29] Table 4 shows which value falls under which category in the confusion matrix from the perspective of the 'Biceps Curl'. For each class, the contents of Table 4 are determined in the same way.

		Actual Class	
		Biceps Curl	**Other Exercise**
	Biceps Curl	TP	FP
Predicted Class	**Other Exercise**	FN	TN

Table 4: Confusion matrix from the perspective of the class 'Biceps Curl'.

TP is the number of predictions where the classifier correctly predicts Biceps Curl as Biceps Curl.
TN is the number of predictions where the classifier correctly predicts Other Exercise as Other Exercise.
FP is the number of predictions where the classifier incorrectly predicts Biceps Curl as Other Exercise.
FN is the number of predictions where the classifier incorrectly predicts Other Exercise as Biceps Curl

The process is illustrated for one class but applies to all classes. Accuracy can then be calculated for each class by using:

$$Class\ Accuracy = (TP + TN)\ /\ (TP + TN + FP + FN)$$

Finally, the average of all class accuracies is resulting in a complete model's accuracy.

5 Results

This section shows the results of the conducted experiments. For each experiment, we will look back at the research questions and describe our findings.

5.1 Classifying Strength Exercises

In this experiment, we look at the research question "To which extent does a neural network perform better in classifying strength exercises than more traditional machine learning techniques?" For the performance assessment of the different models in predicting strength exercises, we use the complete data set. We will start by looking at the performance of the baseline models, after which the setup and performance of the neural network will be explained.

The performance of the traditional machine learning methods models is visible in Table 5. The best performing model was the random forest classifier with a noticeable margin over the other methods.

Model	**Accuracy**
Random Forest	0.545
Decision Tree	0.368
KNN	0.305

Table 5: Performance of the baseline models using the average accuracy over all LOSO folds.

The overall confusion matrices are shown in Figures 6, 7 and 8. These visuals show the bias of the model towards the 'Biceps Curl' and 'Internal- / External Rotation'. These two exercises were most often chosen by the participants. Given that the model contains seven different classes, there is a probability of 0.14 (1/7) for the correct class label when random guessing. By looking at the true positives of all classes in Figure 6, we can see that only the 'Biceps Curl' and 'Internal- / External Rotation' are showing results greater than 0.14. Except for these two exercises, all other exercises can therefore be considered as almost unpredictable by the model. A possible reason for this may be that wheelchair users are often limited to arm-only movements while in a wheelchair, resulting in more difficulties for the classifier to distinguish the exercises. The classifier confused the 'Triceps Extension' most of the time for the 'Biceps Curl', possibly since both exercises target an arm muscle. In addition to the similarity in the muscle activation zone, the model may have a bias towards the 'Biceps Curl', because the 'Biceps Curl' has more samples than the 'Triceps Extension'. It looks like the classifier chooses to classify towards the biased movements, based on the higher number of samples.

When looking at the individual accuracies over all folds, as visible in Table 10 in Appendix C, we see that the performance of the model differs extremely from person to person. The type of exercises performed by the participant is again the most determining factor for this difference. When a participant performed 'Biceps Curl' or 'Internal- / External Rotation', the accuracy is automatically higher. In case of the maximum score of 0.908 for participant AA030, we can see that the participant performed the 'Biceps Curl', 'Internal- / External Rotation' and 'Triceps Extension'. If we look at participant AA036 with the lowest accuracy of 0.144, we see that the 'Shoulder- / Chest Press', 'Biceps Curl' and 'Seated Row' were performed. Participant AA036 performed two less popular exercises, which may explain the low score. The fact that the 'Biceps Curl' was performed, which is an exercise which is generally well-classified exercises. A possible reason that the classifier was not accurate for the 'Biceps Curl' for this specific participant, may have to do with a bad execution form of the exercise. When exercises are performed with a strict form, the model can learn to link a more unified movement pattern to a specific exercise. If the movement patterns of an exercise are more unified, the model has less noise and can better distinguish different exercises. So, a good execution form should also benefit the classifier for exercises that target the same muscle group.

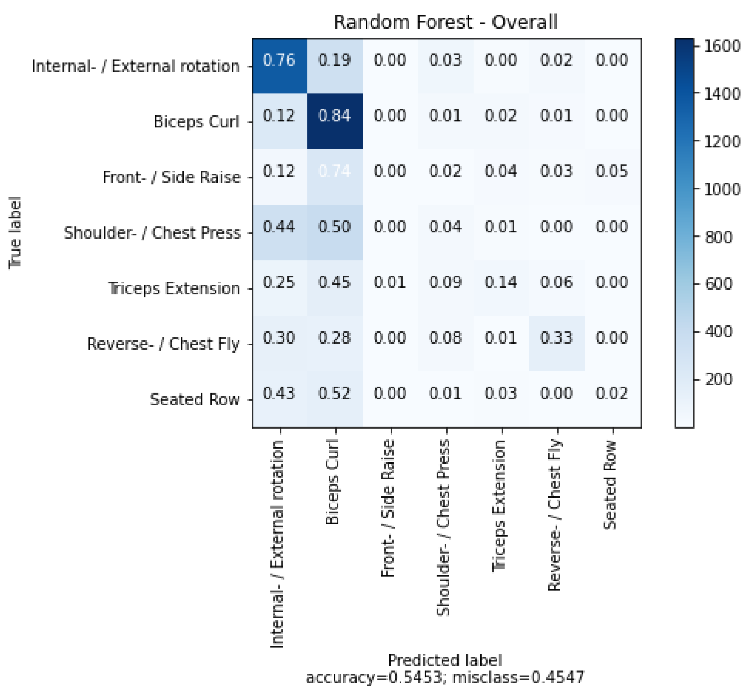

Figure 6: Overall confusion matrix using the Random Forest Classifier.

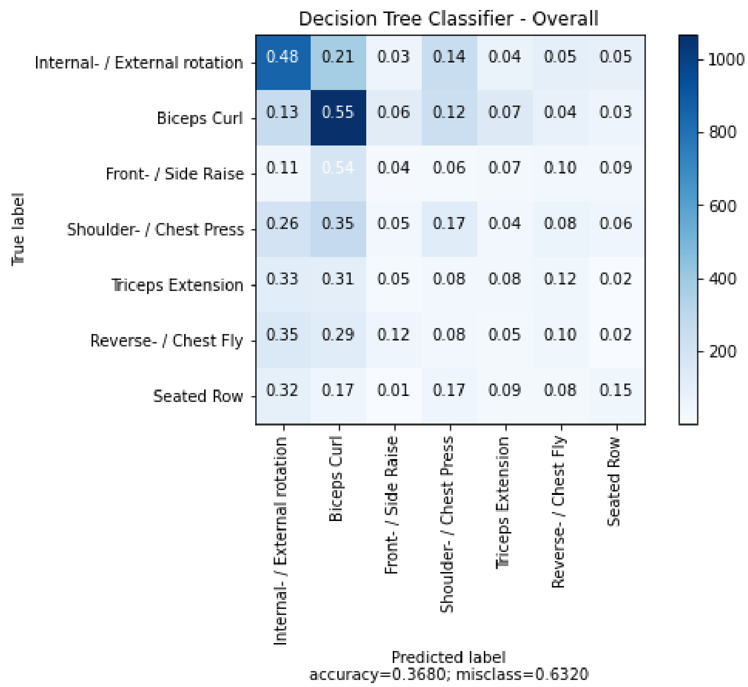

Figure 7: Overall confusion matrix using the Decision Tree Classifier.

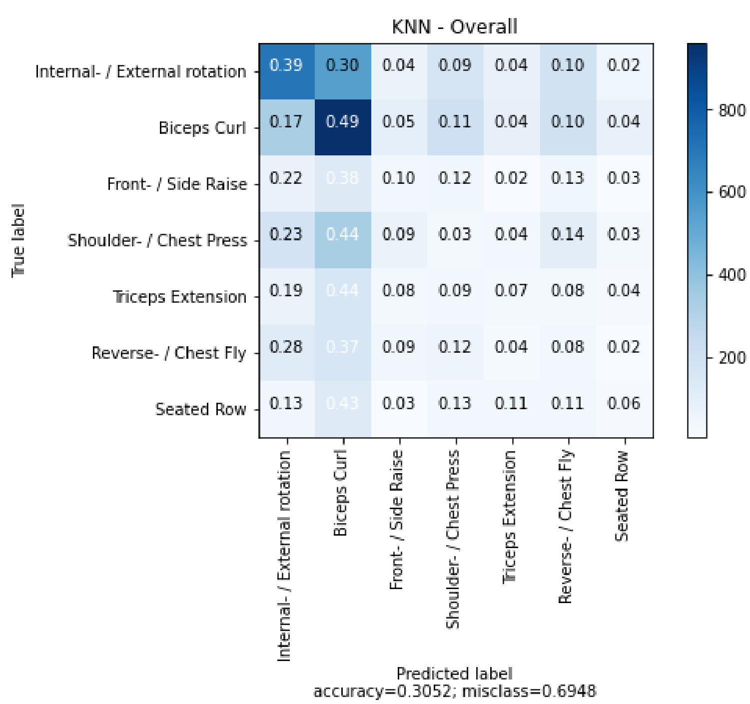

Figure 8: Overall confusion matrix using the KNN Classifier.

For the recurrent neural network (RNN), we need to determine the structure of the network. To find the optimal structure, we randomly selected for each group (SCI, LLA, non-wheelchair user) two random participants. One person belongs to the validation set, and the other person belongs to the test set. The reason for the selection from all wheelchair users groups is so that the model can be tuned on all types of participants and is not tweaked for one specific type of wheelchair user.

Validation set = [AA008, AB003, AC02]
Test set = [AA009, AB002, AC05]

For the parameters of the RNN, as mentioned earlier in section 4.2.2, we first tested the following settings on the validation set. The number of epochs was set to a limited value of 10, so the computational costs are not too high during the tuning process. Besides, we can directly evaluate the performance over the epochs to choose our final amount of epochs. A hidden layer value of 128, was inspired by the HAR model described in the research of Okai et al. [30] Their research tested several different network structures and showed the best performing setting. The parameters as shown below had to be adjusted since it became computational too intensive for our data setup.

- **Amount of hidden layers** [2, 3, 6]
- **Sequence Length** [6, 12, 24]
  Recall that the sampling rate is 12,5 Hz so the sequence length in seconds are respectively [0.5, 1, 2]
- **Batch Size** [32, 64, 128]

We ran each possible combination of the parameters and examined the performance using Tensorboard. For a good performing model, we are looking at a high val_accuracy , as well as a low and stable val_loss.

Hidden Layers	Sequence Length	Batch Size	val_accuracy	val_loss
6	24	32	0.481	1.677
6	12	64	0.411	2.500
2	12	64	0.394	4.128

Table 6: Top 3 performing RNNs based on the validation set.

The top 3 networks based on the val_accuracy, are visible in Table 6. From of the Tensorboard results, we noticed that for most models, the performance drops after 3 epochs, which is why we choose to run LOSO with 3 epochs

After evaluating the best performing models, we had to choose a final for the final performance run using LOSO cross-validation on the entire data set. Running the model using LOSO allows us to compare the accuracy with the baseline models. The best performing network in Table 6 was unfortunately computational too costly to run LOSO, therefore we tested the second and third best performing network using LOSO. The best performing model was eventually the third option from Table 6. This led to the final model, containing two hidden LSTM layers as shown in Figure 9. Dropout layers were added after each hidden layer, to prevent the model from overfitting.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_3 (LSTM)                (None, 12, 32)            8448
_________________________________________________________________
dropout_3 (Dropout)          (None, 12, 32)            0
_________________________________________________________________
lstm_4 (LSTM)                (None, 12, 128)           82432
_________________________________________________________________
dropout_4 (Dropout)          (None, 12, 128)           0
_________________________________________________________________
lstm_5 (LSTM)                (None, 128)               131584
_________________________________________________________________
dropout_5 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 1032
=================================================================
```

Figure 9: Structure summary of the final RNN.

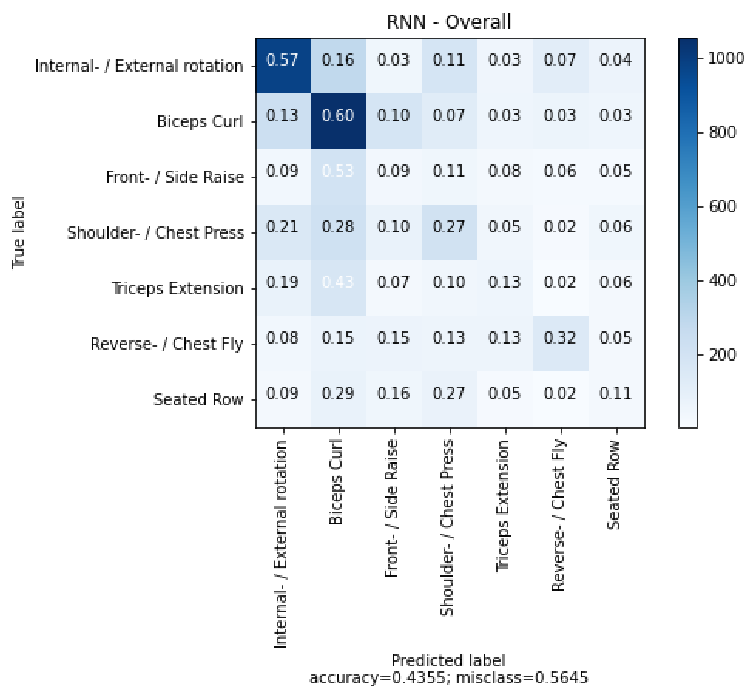

Figure 10: Overall confusion matrix using a Recurrent Neural Network.

The results in Figure 10 show that our RNN performed worse (0.436) than our random forest model (0.545). However, compared to the decision tree and KNN the performance is slightly better. To conclude, our results show that the random forest classifier is the best performing model for classifying strength exercises performed by wheelchair users.

5.2 Comparison Strength Exercises & ADL

In the second experiment, we look at the differences between classifying strength exercises and daily activities (ADL) among wheelchair users. We are diving into the following research question: "What are the differences between predicting strength exercises and daily activities?" To make a fair comparison we use the same models as used for the strength exercises, but then for the data partition of ADL activities. Finally, we compare the results of classifying ADL to the results of classifying strength exercises, as described in the previous section 5.1.

The accuracy scores are illustrated in Table 7. What is immediately noticeable is that the RNN is the best performing model for ADL. Overall the accuracy scores for classifying ADL are comparable to those for strength exercises, despite the higher number of classes.

Model	**Accuracy Strength Exercises**	**Accuracy ADL**
RNN	0.436	0.514
Random Forest	0.545	0.480
Decision Tree	0.368	0.344
KNN	0.305	0.231

Table 7: Accuracy comparison between strength exercises and ADL.

The confusion matrices are visible in Figure 12, 13, 14 and 15. The fact that ADL contains more classes than strength exercises normally results in more difficulties for the classifier. However, as clearly visible in i.e. Figure 12 the model is reasonably consistent in predicting the right classes. The model does not have an extreme bias towards one or two classes, which was the case for classifying strength exercises. The main reason for the better consistency is because the ADL classes are more balanced than the strength classes, as shown in Figure 11.

By inspecting the confusion matrices in Figure 12 and 13, we can see that Adl3, Adl5, Adl8 and Adl10 are the most dominant and best-predicted activities. Recall that the content of the activities are respectively: moving the wheelchair forward at high speed, handing over objects to a person, using a laptop and playing wheelchair basketball. (Table 3). We immediately see that it concerns activities with completely different directions of movement intensity. Because the aforementioned ADL activities are so different in both intensity and direction of movement, these are easier to predict than strength exercises. Besides, the execution of an ADL activity is more unified and clear compared to the execution form of a strength exercise. Wheelchair users are generally more familiar with performing daily activities, as opposed to strength exercises where they sometimes have limited experience.

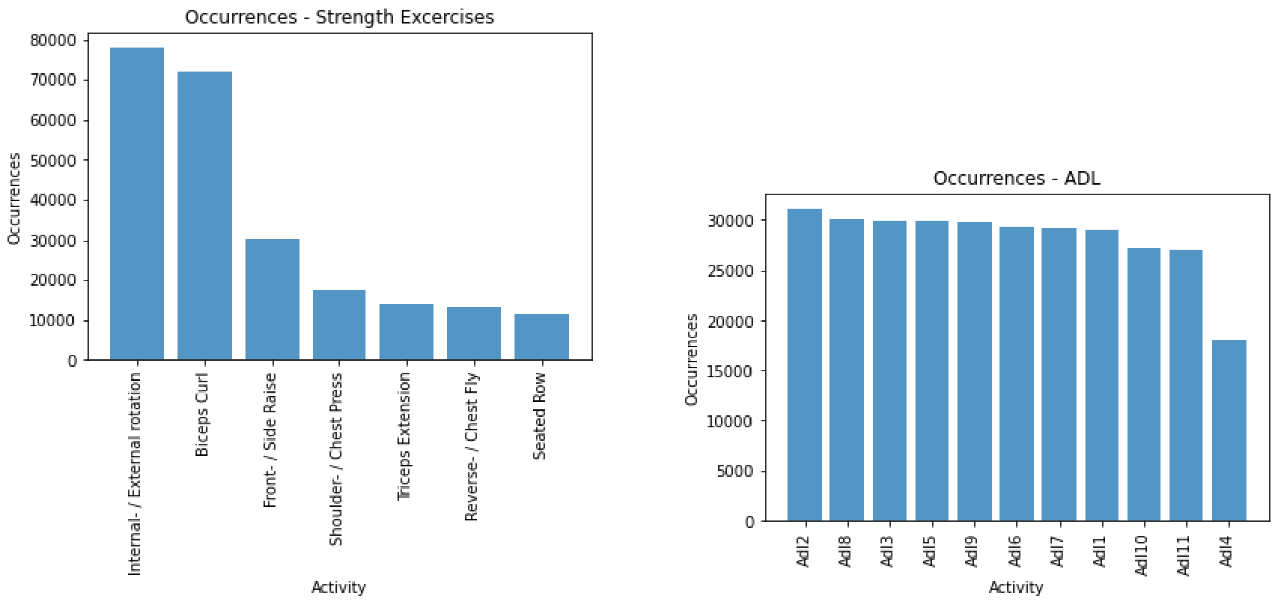

Figure 11: Comparison of class occurrences between Strength Exercises and ADL.

The performance gap between our RNN and traditional machine learning techniques do not show any extreme differences. Looking at the study from Kahn et al., we see that their neural network performs between 0.45 to 0.87 for daily activity classification tasks including non-wheelchair users. [20] Where our model's accuracy is at best 0.514 for classifying ADL, Compared to Kahn et al, the overall accuracy of our models turned out to be at the same level as their worst-performing models. Popp et al. managed to get an accuracy of 0.956 with the KNN classifier. [35] In contrast, our experiment the KNN classifier was the worst classifier with a score of 0.23. Somehow as visible in Figure 15, our KNN model has a strange bias towards Adl1, which is moving the wheelchair forward at slow speed. An explanation for the higher score may the number of accelerometer devices used. Our research used only two wrist sensors and one wheel accelerometer sensor as discussed in section . In comparison, Popp et al. were using two wrist, two ankle, one hip and one chest sensors. [35] Another reason might be the fact that they used more extracted features. Kahn et al. showed a variance of accuracy between 0.45 to 0.87, by using four different feature sets. [20]

To conclude, the main difference between classifying strength exercises and ADL for wheelchair users is the fact that strength exercises are more similar in terms of movement. Because of this similarity, the execution form of an exercise is therefore especially influential in comparison to ADL. Each ADL has a unique movement pattern, making the activity better distinguishable for the classifier.

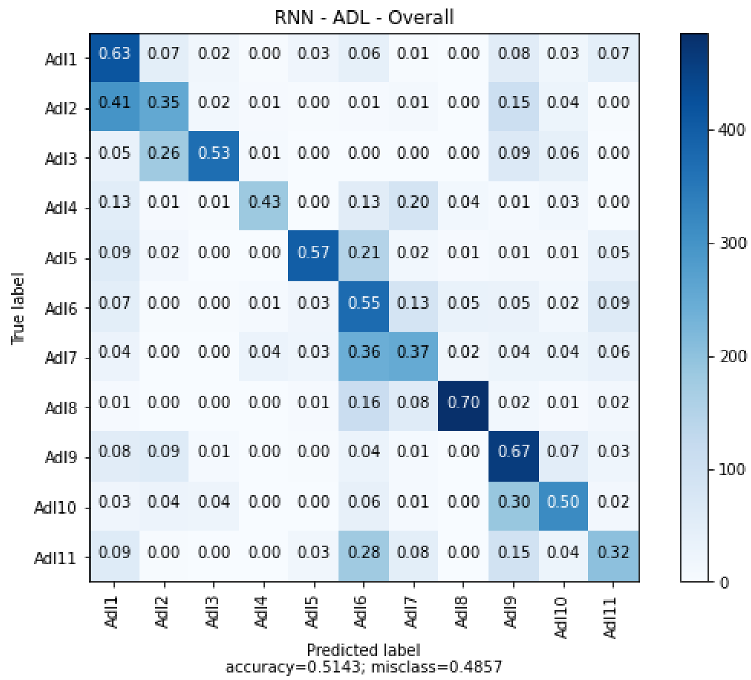

Figure 12: Overall confusion matrix using the Recurrent Neural Network.

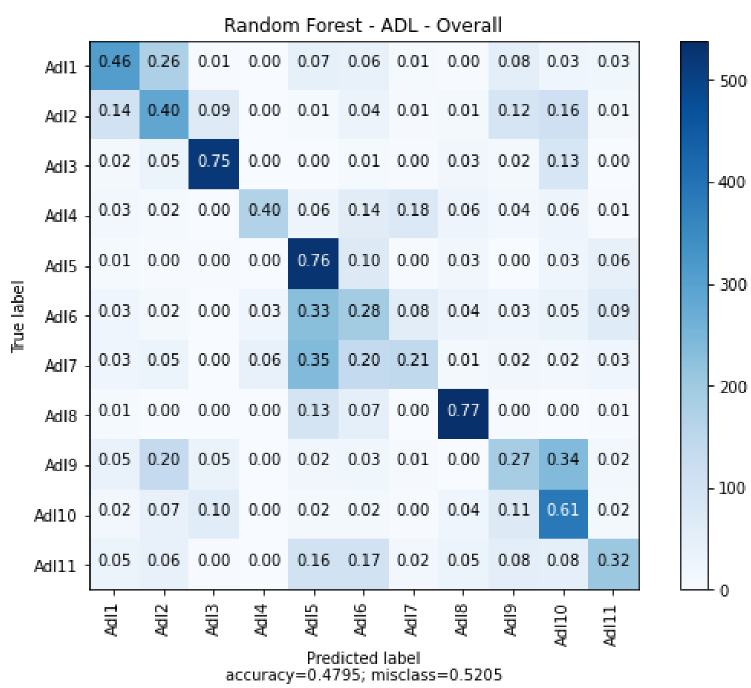

Figure 13: Overall confusion matrix using the Random Forest Classifier.

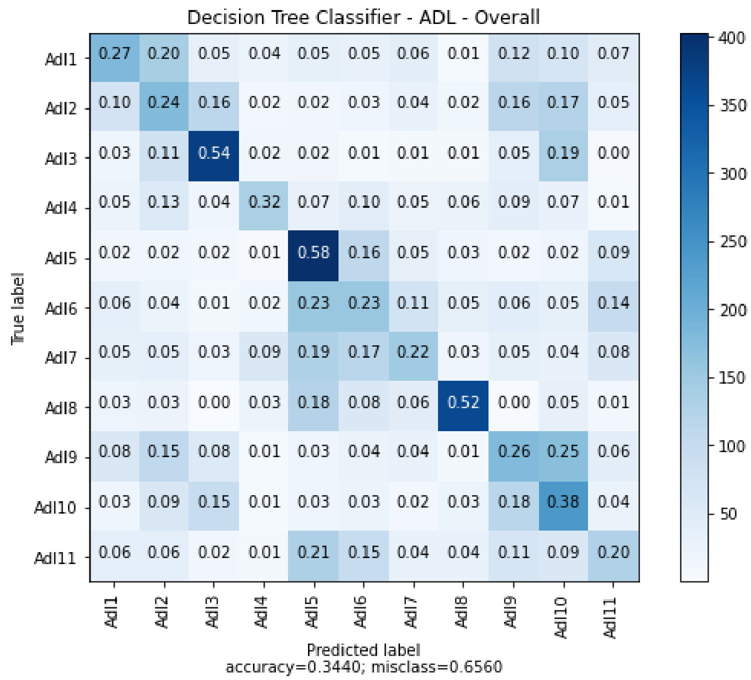

Figure 14: Overall confusion matrix using the Decision Tree Classifier.

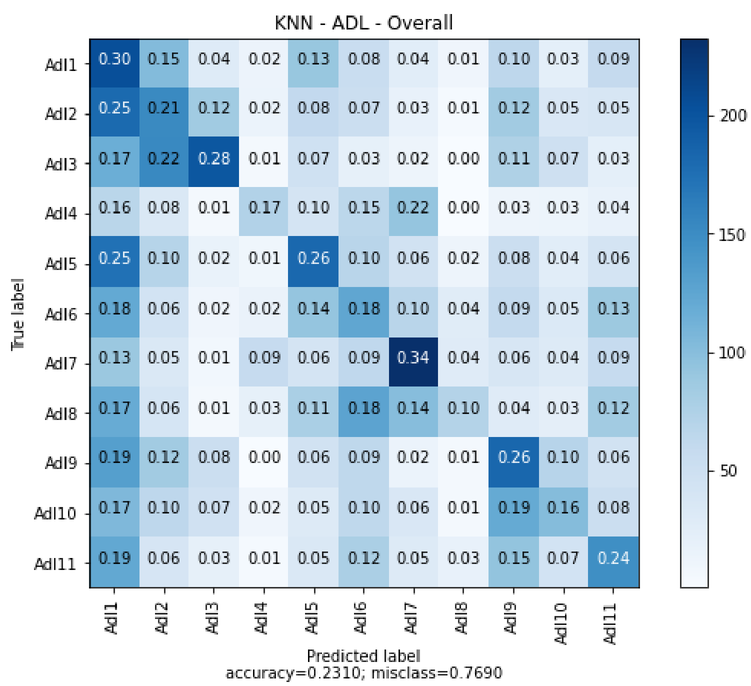

Figure 15: Overall confusion matrix using the KNN Classifier.

5.3 Comparison Types of Wheelchair Users

In the last experiment, we ran the model by using only subsets of the population, based on the injury type. By taking specific subsets, we are trying to find an answer to the last research question: "What are the performance differences when looking at the different types of wheelchair users?" The reason for conducting this experiment is to see if the models are more accurate when there is less variation in the types of wheelchair users. Every wheelchair user can have different limitations even though they share the same injury. The variation limitation within a certain injury group can also be different than other injury groups.

By splitting the data based on injury type, the model also has less data to work with. Recall from Figure 3.1, that the population of the SCI group exists out of 25 participants, the LLA group contains 10 participants and the non-wheelchair user population counts only 5 participants. Especially the non-wheelchair user group is so small that the model can give distorted results. The results are shown in Table 8. We immediately notice that the complete data set generally performs better than injury-specific subgroups, except for the subgroup of LLA wheelchair users. Looking at the different models, we can see that all models performed better when including only LLA wheelchair users.

Model	**Complete**	**SCI**	**LLA**	**NON**
RNN	0.436	0.419	0.530	0.444
Random Forest	0.545	0.502	0.544	0.304
Decision Tree	0.386	0.357	0.439	0.348
KNN	0.305	0.309	0.328	0.266

Table 8: Accuracy comparison for strength exercises between injury-specific sub populations.

From the results, we can see that the model is more accurate when the population only exists out of LLA wheelchair users. This may indicate that this specific group of wheelchair users are more similar to each other regarding their movement limitations. Participants with an LLA injury have the resemblance that they all lack one or more lower limbs of the body. In addition, LLA participants often have functionality in the upper body to a relatively equal extent. When we look at SCI participants, their restrictions on the movement of the upper body can be extremely different from person to person. The spinal cord is nervously connected to all limbs, including the upper body. Therefore, SCI participants can show varying upper body limitations. The other subgroups perform worse than the complete group, making it more logical to use the complete data set. For a machine learning model, it is generally beneficial to include more data, since processing more data leads to better recognition of the underlying connections.

To conclude, recall that a better performing classification model results in a better energy estimation. Because LLA wheelchair users show more similarity as a population, they have a better predictable movement pattern leading to a better performance of the classifier. In the end, there is more potential for a predictable energy estimation for LLA wheelchair users than other participants.

6 Conclusion & Discussion

This last section will describe the conclusion of the research, after which a discussion and suggestions for future research will be given.

6.1 Conclusion

The main research question was: "How can we build a classifier of strength exercises performed by wheelchair users based on accelerometer and heart rate data." We used supervised machine learning to create four different HAR models including seven different exercise categories. Our first experiment showed that classifying strength exercises using the random forest model was more accurate than using the RNN. The accuracy for the random forest classifier was 0.55 compared to an accuracy of 0.44 using the RNN. The overall performance of the model is more accurate than a random guess which equals an accuracy score of 0.14. However, our classification model showed a major bias towards the exercises with the most samples. This was caused by the fact that exercises were free to choose and therefore leading to imbalanced classes.

We also looked at the difference in performance of the model between daily activities exercises and strength exercises. Although the difference in accuracy is not extreme, the ADL model performed more consistently compared to the model for strength exercises. This comparison, strengthens the observation that it is more difficult to classify within the spectrum of fitness exercises. However, we can imagine that fitness as a category can be better distinguished from other activities.

In our last experiment we compared the differences between specific types of wheelchair users. The results showed that LLA wheelchair users are more similar to other wheelchair users. We learned that more similarity within the population paired with more variety within the performed exercise types are beneficial for the classifier.

Finally, we can conclude that our model is currently not performing accurately enough to make a usable classification of strength exercises among wheelchair users. Therefore a performance improvement is needed before the model actually might be useful for an energy estimation model. Consequently, our model is not fitted for deployment to wearables and e-platforms, making it difficult to help wheelchair users in the pursuit for a healthier lifestyle with our model.

6.2 Discussion

After conducting the experiments, we noticed that the models turned out to be weaker than similar classification models from Popp et al. and Okai et al. [35, 30] Class imbalance seems to be the main reason for the worse performance. Participants were free to choose the strength exercises, leading to a wide variety of exercise types. Even after the normalization of the researcher's interpretation of exercise naming, the number of unique exercises remained fourteen. Certain exercises were only performed by a few participants, making it very hard for the classifier to learn from the low number of samples. These exercises were renamed and regrouped, leading to the influence of our interpretation of the exercises. Some people may categorize a certain exercise in a different exercise group, consequently leading to different results. Also, the regrouping may blur the boundaries between exercises. In our experiments regrouping the exercises contributed to better accuracy, but regrouping has also a downside. When overdone, each regrouped category can contain so many diverse movement patterns that the classes start to look alike, resulting in lower accuracy.

A common way to handle class imbalance is by using artificial resampling. Imblearn is a package that enables this kind of resampling techniques. In this approach, a method called SMOTE artificially upsamples all minority classes to match the level of the majority class. In this study, we also experimented with artificial resampling using SMOTE. However, the results were not as expected and in our case, the model performed worse than without resampling. In the next section, we will give suggestions which are interesting to look at in future research.

6.3 Future Research

A suggestion to extend this research and possibly improved the model is by collecting data using a predetermined selection of exercises. This way the classes are more balanced, which is beneficial for the model as shown by our experiment in section 5.2. The ADL activities were predetermined and therefore have a relatively equal number of samples as shown in Figure 11. This led to a more consistent model. By predetermining the exercises, interpretation of exercises will likely play a smaller role and the researcher can focus more on monitoring the correct execution of the exercise. Of course, a lab controlled setting also has potential drawbacks. For example, predetermining exercises will lead to cases where a wheelchair user is unable to perform a particular exercise. The degree of limitation of movement among wheelchair users varies from person to person, resulting in possible hiccups while collecting data. In addition, a controlled lab setting may make the model less compatible with deployment in a real-world setting. In a real-world setting, there is also a free choice of exercises.

Another approach to improve the accuracy of the model is to add more features, including more demographic information about the participants. Since every individual is different in athletic capacity, the rate of perceived exertion (RPE) indicates the intensity of an exercise. The intensity of each set was rated by the participant on a scale from one to ten. By combining the exercise weight and corresponding RPE, an estimation about a person's capacity can be determined. [11] Due to the variance of limitation in movement, the execution form of an exercise can differ per wheelchair user. When demographic and capacity-related features are added, the model will rely less on purely accelerometer data. This reduces the effect of different execution forms of exercises. In addition, it could be interesting to add more sensors, which gives more information for the model about the movement of other body parts.

References

[1] M. M. Adams and A. L. Hicks. Spasticity after spinal cord injury. *Spinal Cord*, 43(10):577, 2005. Accessed: June 5, 2020.

[2] A. J. T. Bakkum, S. D. Groot, J. M. Stolwijk-Swüste, D. J. V. Kuppevelt, L. H. V. V. D. Woude, and T. W. J. Janssen. Effects of hybrid cycling versus handcycling on wheelchair-specific fitness and physical activity in people with long-term spinal cord injury: a 16-week randomized controlled trial. *Spinal Cord*, 53(5), 2015. Accessed: July 12, 2020.

[3] S. Benedetto, C. Caldato, E. Bazzan, D. Greenwood, V. Pensabene, and P. Actis. Assessment of the fitbit charge 2 for monitoring heart rate. 2018. Accessed: February 27, 2020.

[4] M. Berchtold, M. Budde, D. Gordon, H. R. Schmidtke, and M. Beigl. Actiserv: Activity recognition service for mobile phones. In *International Symposium on Wearable Computers (ISWC) 2010*, pages 1–8, 2010. Accessed: June 22, 2020.

[5] D. M. Bravata, C. Smith-Spangler, V. Sundaram, A. L. Gienger, N. Lin, R. Lewis, C. D. Stave, I. Olkin, and J. R. Sirard. Using Pedometers to Increase Physical Activity and Improve HealthA Systematic Review. *JAMA*, 298(19):2296–2304, 11 2007. Accessed: June 22, 2020.

[6] R. Cachucho, M. Meeng, U. Vespier, S. Nijssen, and A. Knobbe. Mining multivariate time series with mixed sampling rates. 2014. Accessed: April 12, 2020.

[7] P. Carrington, K. Chang, H. Mentis, and A. Hurst. ”but, i don’t take steps”: Examining the inaccessibility of fitness trackers for wheelchair athletes. pages 193–201, 10 2015. Accessed: June 22, 2020.

[8] A. Chakure. K-nearest neighbor (k-nn)?, 2019. https://towardsdatascience.com/decision-tree-classification-de64fc4d5aac. Accessed: June 2, 2020.

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Gated feedback recurrent neural networks. 2015. Accessed: June 15, 2020.

[10] E. A. Claridge, R. J. G. van Den Berg-Emons, H. L. D. Horemans, W. M. A. van Der Slot, N. van Der Stam, A. Tang, B. W. Timmons, J. W. Gorter, and J. B. J. Bussmann. Detection of body postures and movements in ambulatory adults with cerebral palsy: a novel and valid measure of physical behaviour. *Journal of NeuroEngineering and Rehabilitation*, 16(1):1–11, 2019. Accessed: February 27, 2020.

[11] Cleveland Clinic. Rated perceived exertion (rpe) scale. https://my.clevelandclinic.org/health/articles/17450-rated-perceived-exertion-rpe-scale Accessed: June 5, 2020.

[12] M. Fanchamps, H. Horemans, G. Ribbers, H. Stam, and J. Bussmann. The accuracy of the detection of body postures and movements using a physical activity monitor in people after a stroke. *Sensors*, 18(7):2167, 2018. Accessed: February 27, 2020.

[13] FAPESP and NWO. Application form - data-driven research on sports & healthy living, 2016. Accessed: February 18, 2020.

[14] E. Füzéki, T. Engeroff, and W. Banzer. Health benefits of light-intensity physical activity: A systematic review of accelerometer data of the national health and nutrition examination survey (nhanes). *Sports Medicine*, 47(9):1769–1793, 2017. Accessed: February 18, 2020.

[15] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1999. Accessed: June 30, 2020.

[16] A. P. Hills, N. Mokhtar, and N. M. Byrne. Assessment of physical activity and energy expenditure: An overview of objective measures. *Frontiers in Nutrition*, 1:5, 2014. Accessed: February 20, 2020.

[17] O. Horgan and M. Maclachlan. Psychosocial adjustment to lower-limb amputation: A review. *Disability and Rehabilitation*, 26(14-15):837–850, 2004. Accessed: June 8, 2020.

[18] L. P. Hua, C. A. Brown, S. J. Hains, M. Godwin, and J. L. Parlow. Effects of low-intensity exercise conditioning on blood pressure, heart rate, and autonomic modulation of heart rate in men and women with hypertension. *Biological Research For Nursing*, 11(2):129–143, 2009. Accessed: February 27, 2020.

[19] T. W. Janssen and D. Hoevenaars. C1 onderzoeksprotocol energy expenditure in rest, daily activities and during strength activities in individuals with spinal cord injury or lower limb amputation, May 2018. Accessed: March 3, 2020.

[20] A. M. Khan, M. H. Siddiqi, and S.-W. Lee. Exploratory data analysis of acceleration signals to select light-weight and accurate features for real-time activity recognition on smartphones. *Sensors (Basel, Switzerland)*, 13(10):13099–13122, 2013. Accessed: June 5, 2020.

[21] P. Kokkinos. Physical activity, health benefits, and mortality risk. *ISRN Cardiology*, 2012, 2012. Accessed: March 15, 2020.

[22] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys Tutorials*, 15(3):1192–1209, 2013. Accessed: June 22, 2020.

[23] A. J. Littman, E. J. Boyko, M. L. Thompson, J. K. Haselkorn, B. J. Sangeorzan, and D. E. Arterburn. Physical activity barriers and enablers in older veterans with lower-limb amputation. *Journal of rehabilitation research and development*, 51(6):895–906, 2014. Accessed: June 8, 2020.

[24] A. Mannini and A. M. Sabatini. Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors (Basel, Switzerland)*, 10(2):1154–1175, 2010. Accessed: June 6, 2020.

[25] A. Mannini, L., M. Rosenberger, M., W. Haskell, S., A. Sabatini, S., and S. Intille, S. Activity recognition in youth using single accelerometer placed at wrist or ankle. *Medicine & Science in Sports & Exercise*, 49(4):801–812, 2017. Accessed: June 6, 2020.

[26] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher. Activity recognition and monitoring using multiple sensors on different body positions. pages 4 pp.–116, 2006. Accessed: June 22, 2020.

[27] Mayo Clinic. Spinal cord injury - symptoms causes. https://www.mayoclinic.org/diseases-conditions/spinal-cord-injury/symptoms-causes/syc-20377890. Accessed: June 6, 2020.

[28] L. Meazzini. Everything you need to know about time series, 2020. https://towardsdatascience.com/everything-you-need-to-know-about-time-series-5fa1834d5b18. Accessed: June 22, 2020.

[29] J. Mohajon. Confusion matrix for your multi-class machine learning model, 2020. https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826. Accessed: July 5, 2020.

[30] J. Okai, S. Paraschiakos, M. Beekman, A. Knobbe, and C. Rebelo de Sá. Building robust models for human activity recognition from raw accelerometers data using gated recurrent units and long short term memory neural networks. 2019. Accessed: June 5, 2020.

[31] O.-E. Olsen, G. Myklebust, L. Engebretsen, I. Holme, and R. Bahr. Exercises to prevent lower limb injuries in youth sports: cluster randomised controlled trial. *BMJ*, 330(7489):449, 2005. Accessed: June 30, 2020.

[32] B. K. Pedersen and B. Saltin. Evidence for prescribing exercise as therapy in chronic disease. *Scandinavian journal of medicine & science in sports*, 16 Suppl 1:3, 2006. Accessed: June 25, 2020.

[33] A. C. Pinheiro Volp, F. C. Esteves de Oliveira, R. Duarte Moreira Alves, E. A. Esteves, and J. Bressan. Energy expenditure: components and evaluation methods. *Nutrición Hospitalaria*, 26(3):430–440, 2011. Accessed: February 20, 2020.

[34] W. L. Popp, L. Richner, M. Brogioli, B. Wilms, C. M. Spengler, A. E. P. Curt, M. L. Starkey, and R. Gassert. Estimation of energy expenditure in wheelchair-bound spinal cord injured individuals using inertial measurement units. *Frontiers in Neurology*, 9:478, 2018.

[35] W. L. Popp, S. Schneider, J. Bär, P. Bösch, C. M. Spengler, R. Gassert, and A. Curt. Wearable sensors in ambulatory individuals with a spinal cord injury: From energy expenditure estimation to activity recommendations. *Frontiers in Neurology*, 10:1092, 2019. Accessed: June 22, 2020.

[36] C. Schaefer, A., C. Nigg, R., J. Hill, O., L. Brink, A., and R. Browning, C. Establishing and evaluating wrist cutpoints for the geneactiv accelerometer in youth. *Medicine & Science in Sports & Exercise*, 46(4):826–833, 2014. Accessed: February 27, 2020.

[37] T. Stöttner. Why data should be normalized before training a neural network, 2019. https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d. Accessed: June 2, 2020.

[38] S. T. Entropy: How decision trees make decisions, 2019. https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8. Accessed: July 4, 2020.

[39] E. M. Tapia, S. S. Intille, W. Haskell, K. Larson, J. Wright, A. King, and R. Friedman. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. pages 37–40, 2007. Accessed: June 22, 2020.

[40] A. E. Tawashy, J. J. Eng, K. H. Lin, P. F. Tang, and C. Hung. Physical activity is related to lower levels of pain, fatigue and depression in individuals with spinal-cord injury: a correlational study. *Spinal Cord*, 47(4):301, 2008. Accessed: June 22, 2020.

[41] Techopedia. K-nearest neighbor (k-nn)?, 2017. https://www.techopedia.com/definition/32066/k-nearest-neighbor-k-nn. Accessed: June 2, 2020.

[42] R. J. van Den Berg-Emons, J. B. Bussmann, and H. J. Stam. Accelerometry-based activity spectrum in persons with chronic physical conditions. *Archives of Physical Medicine and Rehabilitation*, 91(12):1856–1861, 2010. Accessed: June 25, 2020.

[43] Virtuagym Nederland. Virtuagym, 2020. https://business.virtuagym.com/, Accessed: February 19, 2020.

[44] T. Yiu. Understanding random forest, 2019. https://towardsdatascience.com/understanding-random-forest-58381e0602d2. Accessed: June 2, 2020.

[45] Z. Zhang, Y. Song, L. Cui, X. Liu, and T. Zhu. Emotion recognition based on customized smart bracelet with built-in accelerometer. *PeerJ*, 4(7), 2016. Accessed: June 22, 2020.

Appendices

A Added Features Overview

This section lists and describes all added features during the research.

Feature	Description
MeanX	Mean of X-axis signal from the GENactiv, 8s time window.
StdX	Standard Deviation of X-axis singal from the GENactiv, 8s time window.
MinX	Min of X-axis signal from the GENactiv, 8s time window.
MaxX	Max of X-axis signal from the GENactiv, 8s time window.
MeanY	Mean of Y-axis signal from the GENactiv, 8s time window.
StdY	Standard Deviation of Y-axis signal from the GENactiv, 8s time window.
MinY	Min of Y-axis signal from the GENactiv, 8s time window.
MaxY	Max of Y-axis signal from the GENactiv, 8s time window.
MeanZ	Mean of Z-axis signal from the GENactiv, 8s time window.
StdZ	Standard Deviation of Z-axis signal from the GENactiv, 8s time window.
MinZ	Min of Z-axis signal from the GENactiv, 8s time window.
MaxZ	Max of Z-axis signal from the GENactiv, 8s time window.
MeanHeartrate	Average heart rate from the Fitbit, 8s time window.
Magnitude_GEN_acc	Magnitude of the GENactiv signal $\sqrt{(x^2 + y^2 + z^2)}$
Magnitude_ACT8RWrXYZ	Magnitude of the Activ8 (Wrist) signal $\sqrt{(x^2 + y^2 + z^2)}$
Magnitude_ACT8RWhXYZ	Magnitude of the Activ8 (Wheel) signal $\sqrt{(x^2 + y^2 + z^2)}$
Velocity_Gen_acc	Velocity of the GENactiv signal
Velocity_ACT8RWrXYZ	Velocity of the Activ8 (Wrist) signal
Velocity_ACT8RWhXYZ	Velocity of the Activ8 (Wheel) signal

Table 9: Features added to the existing raw features.

B Feature Correlation Matrix

This section shows the correlation between all features.

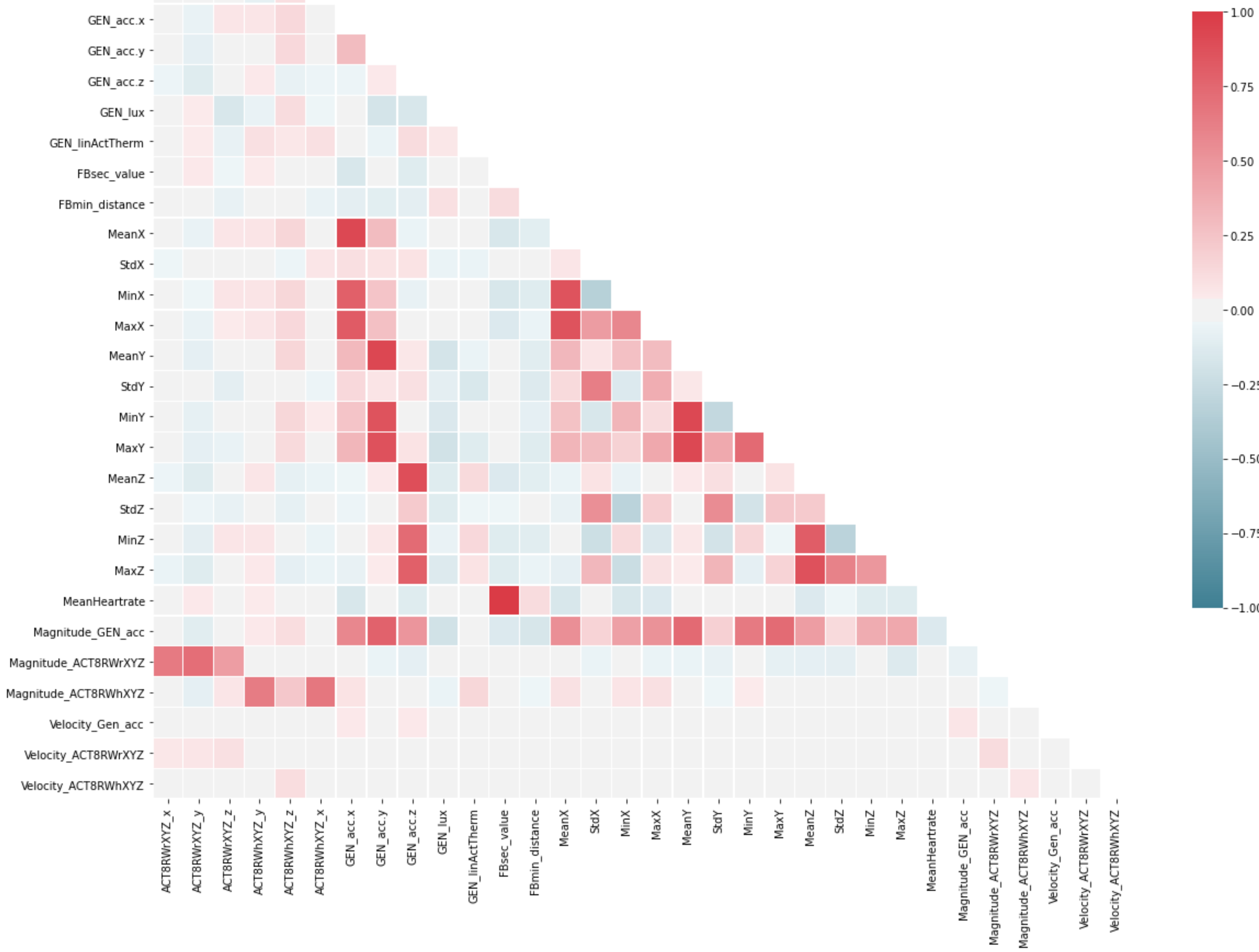

Figure 16: Correlation matrix of the features.

C Accuracy Variation

This section shows the variation in accuracy across the different participants. The illustrated scores are obtained using the random forest classifier running LOSO on the complete data set.

Fold	**Accuracy**
1	0.3883462445267767
2	0.3407441716844218
3	0.6492447533752173
4	0.572594501718213
5	0.3570072161033042
6	0.7615696887686062
7	0.5706438145462536
8	0.3401462994836489
9	0.37769971376528755
10	0.6430847673677501
11	0.7838574804580986
12	0.6298551678736011
13	0.4890846922672278
14	0.794452347083926
15	0.5870646766169154
16	0.7265625
17	0.731631679389313
18	0.6778485491861288
19	0.5088097806544408
20	0.9082429501084599
21	0.43051335359027665
22	0.36806993328732457
23	0.5299810645907848
24	0.14434503939110696
25	0.38708414872798436
26	0.354870974194839
27	0.5409784979744469
28	0.5021147014041617
29	0.5390990225244369
30	0.7068348389296063
31	0.42314572726021743
32	0.635669276303227
33	0.7528455284552845
34	0.7931778929188256
35	0.38639852625115134
36	0.7192892533743379
37	0.4454291044776119
38	0.5227226411249236
39	0.5487155388471178
40	0.48343232260081276

Table 10: Variation in accuracy over all folds using the Random Forest Model.