Predicting the well-being of COPD patients with respiratory data from pulmonary rehabilitation

Darius Fischer

Master of Science

School of Informatics
University of Edinburgh

2016

Abstract

Chronic obstructive pulmonary disease (COPD) is a long-term lung disease which severely reduces the quality of life of the affected patients. Timely changes to the patient's medication in case of acute worsening of the condition (*exacerbation*) are essential for successful treatment. However, these changes in condition are often left unreported until hospital admission is inevitable. In this thesis, a new, automatic approach to assess the well-being of COPD based on the respiratory signals recorded with a three-axis accelerometer (*RESpeck*) is evaluated. Thirty-one subjects with COPD completed pulmonary rehabilitation over a period of 4.5 months while wearing the RESpeck and regularly filled out the *COPD Assessment Test (CAT)* which scored their subjective well-being on a scale of 0 to 40 points. With features extracted from the RESpeck data, the CAT score could be predicted within a mean absolute error of 2.77 points, and with a root-mean-square error of 4.639.

Acknowledgements

I would like to express my sincere gratitude towards my supervisor, DK Arvind, for his continuous support at all stages of the project, for his insightful remarks and his guidance in phrasing complicated topics in a simple way.

Additionally, I am grateful for the work of his research assistant, Andrew Bates, who provided me with the raw study data, helped me with any technical questions regarding the RESpeck sensor and its output, and integrated my questionnaire code into the *RESpeck Rehab* Android application.

I am also thankful for countless insightful conversations with my fellow students regarding general machine learning techniques.

Lastly, my whole study and this thesis would not have been possible without the support of my family and friends.

Table of Contents

Chapter 1

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a lung disease commonly affecting long-term smokers, and people who were regularly exposed to fine particulate pollution such as miners. It is estimated to be the fifth highest cause of death worldwide, and is projected to move up to the fourth place by 2030 [35]. The airways of COPD patients are chronically inflamed and blocked with mucus, impeding expiration. This means that the oxygen supply to the body is limited and affected patients are exhausted faster from normal day-to-day activities than healthy people which can severely affect their quality of life [21].

The most important remedy for COPD patients is to quit smoking, followed by physical exercises (e.g. pulmonary rehabilitation) and medication. The latter two treatment options depend on individual factors such as the medical record of the patient, their age and the short-term development of the disease.

An "acute [...] worsening of the patients respiratory symptoms that is beyond normal day-to-day variations" [21] is called an *exacerbation*. Exacerbations are the leading factor of mortality within COPD patients and also constitute the most common reason for hospital admissions in the UK [43, 48]. It is therefore of great interest to detect changes in the patient's health as soon as possible, so that the medication can be adjusted and hospital treatment avoided.

However, patients often do not report health changes to their physicians [28] which can be attributed to a mixture of playing down symptoms and adapting to new health issues quickly. In addition, subjective assessments will only cover relative changes to a patient's health and not allow different patients do be compared.

For this reason, tests like the FEV_1 score which measures the maximum amount of air a patient can exhale in one second have been developed. This test allows physicians to categorise patients into different stages of COPD, and also confirm exacerbations. However, this test is taken at specific points in time rather than continuously which means it still relies the active decision of the patient to record their condition.

It is therefore desirable to have a continuous monitoring and alert system which detects changes in the subject's health while they are at home. The *RESpeck* device is an accelerometer sensor which is worn as a plaster on the patients chest and was developed for measuring the respiratory rate in a simple and unobtrusive way. It has already been validated as a respiratory monitor in previous studies [16].

The data being analysed is taken from subjects with COPD who wear the RESpeck when performing their pulmonary rehabilitation exercises. It comprises motion data during exercises and rest periods. The subjects regularly filled out a questionnaire widely used in clinical practice called the *COPD Assessment Test (CAT)* [27] which outputs a subjective score of well-being between 0-40 points. The hypothesis is that the breathing signal extracted from the RESpeck during pulmonary rehabilitation can be used to assess the subject's health as measured by the CAT.

In this work, we will first develop and validate a new method for calculating the breathing signal from the acceleration data. From this signal, different time-domain and frequency-domain features can then be extracted which are either based on previous research about breathing disorder classification (e.g. [17]), or based on intuition and validated empirically. To our best knowledge, this is the first work trying to predict the general well-being of COPD patients from continuous respiratory data, rather than only detect very specific breathing patterns. Even though many ideas are taken from previous research, most of the methods and extraction techniques are original.

In total, 434 features are extracted from 270 data samples. We evaluate the predictive capability with linear regression models and Artificial Neural Networks and are able to reduce the prediction error down to a root-mean-square error of 4.639 or mean absolute error of 2.77 (on a scale of 0-40 points) on a separate test set, which is considerably better than the minimum baseline of 12.03 (RMSE) or 11.34 (MAE). Additionally, the correlation coefficients of features and CAT scores contain clinically relevant insights which are discussed in detail.

In the rest of this thesis, Chapter 2 outlines the related literature and Chapter 3 provides background information which sets this work in context. Chapter 4 describes the labels and features in detail which are used for the regression task, and Chapter 5 investigates whether the main hypothesis is justified and which features discriminate the best between healthy and unhealthy subjects. Lastly, Chapter 6 draws conclusions for this study and provides insights into future work that can be undertaken.

Chapter 2

Related work

This chapter begins with an overview on the devices used for respiratory monitoring and introduces the *RESpeck* sensor which was deployed in the studies of this thesis. This is followed by a review on current methods for assessing the well-being of COPD patients. Lastly, different features and techniques for detecting and classifying breathing disorders are discussed.

2.1 Devices for respiratory monitoring

There are many approaches for measuring the breathing signal of patients. In clinical practice, spirometers are used to detect restricted breathing with short-term measurements of only a few seconds [11, 21], while nasal cannulas are used in polysomnographies (sleep studies) for longer recordings (cf. [25]). If only the respiratory rate is of interest, it can also be visually observed by a nurse [5].

Beside the pressure sensors which directly measure in- and exhaled air, there are other, less costly and obtrusive devices to record the respiratory signal. Thoracic or abdominal belts pick up the expansion of the thorax or abdomen [14], acoustic sensors extract the breathing signal from the breathing sounds of a patient in otherwise quiet environments [13] and accelerometer sensors record the acceleration of the thorax or abdomen. The latter sensor type provides a good balance between accuracy of the calculated breathing signal and practicality. Bates et al. [16] successfully validated an acceleration sensor called *Orient speck* in a study on patients having received opioid analgesia after surgery.

They were able to match the acceleration-based breathing signal with the signal from a nasal cannula within 2 bpms (breaths per minute) in 86% of occasions.

The successor of the Orient sensor, called *RESpeck* (see Figure 2.1), was used in the studies for this thesis. Sensors for respiratory monitoring have to our best knowledge exclusively been deployed in clinical or laboratory settings, often with healthy subjects (cf. [14, 18]). One reason is the complexity of most of these sensors which makes them unsuitable to be deployed without the presence of a researcher or doctor. Another is that most of the sensors are too obtrusive to be worn for a longer period of time without restricting the subject in their daily lives. The RESpeck, however, fastened with a plaster just below the rib cage, is almost unnoticeable. To our best knowledge, the RESpeck is the first sensor used for recording the respiratory rate at the patient's home.

Figure 2.1: The RESpeck sensor in comparison to a 50 pence coin.

2.2 Assessing the well-being of COPD patients

The *Global Initiative for Chronic Obstructive Lung Disease (GOLD)* is an international organisation publishing regularly updated guidelines about the assessment and treatment of COPD [21]. They recommend the FEV_1/FVC metric as the best indicator for the severity of COPD. This metric captures the amount of air a patient can exhale in a second (forced expiratory volume), divided by the total volume which he can exhale (forced vital capacity). If this ratio falls below 0.7, and the patient additionally shows the typical symptoms for COPD (breathlessness, chronic coughing, sputum production, chest tightness), the patient should be diagnosed with COPD. After the diagnosis, it is sufficient to record the FEV_1 score on its own without the additional FVC [21] to measure changes over time.

Other metrics for dividing the patients into different categories of severity is the BODE index or the SCOPEX index. The BODE index extends the FEV_1 score with three other metrics: the body mass index (BMI), the prevalence of dyspnea (i.e. breathlessness) as measured by a questionnaire, and the distance a person can walk in six minutes [34]. SCOPEX adds the sex of the patient and the mean daily reliever use (such as Salbutamol) to the FEV_1 score [32].

SCOPEX was specifically designed to predict the chances of an acute worsening of the patient's condition, termed *exacerbation* [10]. These frequency and intensity of exacerbations are the best predictor for the mortality of COPD patients. They are also the leading cause of hospitalisations in the UK and pose a high economic burden [48]. Foreseeing and preventing exacerbations is therefore the main goal of COPD treatment [21]. Ideally, exacerbations are addressed as soon as possible in order to mitigate their effect. However, it is often difficult for patients to assess when the change of symptoms is high enough to warrant medical attention [28], i.e. when the fluctuations of their well-being is "beyond normal day-to-day variation" [21]. A metric like the FEV_1, taken from a portable spirometer, can objectify the condition and help patients decide whether to see a doctor, albeit this requires the patient to regularly make recordings with such a device. Smith et al. let patients note down different COPD-related metrics daily in a smartphone application [42], such as their breathlessness and sputum quantity. Additionally, the patients wrote down the best of three hand-held spirometer recordings. The information from the device and the questionnaire answers were combined into one score which provided a good estimate of the patient's condition. If this score reached a certain threshold, the application triggered an alert and notified medical staff who can then contact the patient do discuss further medication.

The questions posed by Smith et al. in their application closely resemble those of the *COPD Assessment test* [20], a questionnaire which is prevalent in clinical practice due to its simplicity (only 8 questions) and expressiveness [21, 44]. Other well-known questionnaires are the *Clinical COPD Questionnaire (CCQ)* (10 questions), and the *St George's Respiratory Questionnaire (SGRQ)* (14 questions with additional subquestions). Both the CAT and CCQ have been shown to correlate well with other metrics for assessing COPD, such as the FEV_1 value (as captured by the GOLD index) and the BODE index, while being easier to conduct as they can be self-administered by the patient [27, 36, 44]. The CCQ is targeted towards a completion every week, as the questions specifically ask about the past week, whereas the CAT is formulated in present

tense and can therefore be completed anytime. This makes it easier to detect short-term, "acute" changes. The CAT questionnaire was therefore chosen as the method for assessing the patient's well-being in this project, due to its locality in time, its simplicity, and its close correlation to other metrics used in clinical practice.

2.3 Features for pattern detection in breathing signals

As seen in the previous section, the diagnosis of COPD is currently made from very short breathing recordings through a spirometer and more high-level features such as the breathlessness of the patient. There is, to our best knowledge, so far no single tested feature or pattern extracted from longer breathing signals which allows deductions about the health of the COPD patient. Such a method would have several advantages compared to the current metrics: In longer recordings, errors in the readings are averaged out and the extracted feature values are therefore more stable. Longer recordings also allow features which span a longer period, such as the breathing recovery rate measured over a whole rest period after exercises (see Section 4.2.2.1). Lastly, the extracted features can be multidimensional, i.e., there can be multiple features for a single timestamp. Recordings with the RESpeck are also passive, i.e., does not require any action of the patient apart from initiating the connection to the smartphone, which makes the sensor more likely to be actually used at home.

As there is currently no study on breathing monitoring of COPD patients at home, we instead looked at studies aimed at detecting specific breathing disorders, such as *sleep apnea* (pauses in breathing due to blocked airways), *Bradypnea* or *Tachypnea* (unusually slow or fast breathing). The assumption is that some of these patterns also allow deductions about the state of COPD.

At first, it might seem easier to bypass the feature selection process completely and simply feed the whole breathing signal in fixed windows into a neural network. Varady et al. [46] investigated this approach for the detection of sleep apnea and found it to give poor results in the classification task. They attributed this to the diverse length, amplitude height and general form of breaths, even within the signal of one patient. Furthermore, each window will begin and end at a different part of a breath which makes it difficult for the network to learn specific patterns for each input node. This

kind of analysis is only possible for very short spirometer recordings, where by default each recorded breath has a similar form [3, 47].

Instead of using the raw signal, Varady et al. extracted what they call the *instantaneous respiration amplitude (IRA)* and the *instantaneous respiration interval (IRI)*. These features are stored as their own time-domain signals, but reduced in information compared to the original signal. IRA captures the maximum amplitude of each breath and IRI the interval, i.e. duration of each breath. Using these two features extracted from a nasal airflow signal as input to a neural network, the authors were able to classify apnea and hypopnea with a specificity of 88.7% and 91.0%, and a sensitivity of 97.0% and 78.7% respectively. Adding the features from an additional breathing signal recorded by an abdominal or thoracic excursion belt only improved the performance slightly [46].

Koley et al. [30] similarly devised a system for the detection of apnea and hypopnea events, but additionally deploy it as a real-time alert system. They started with a large set of features, which also contained the IRA and IRI, and selected those most helpful for the task with the *F-score* (not to be confused with the F1-score) and the *Recursive Feature Elimination* technique (details in Koley et al. [31]). They discriminate their features into three different types: time-domain based features, frequency-domain based features and non-linear features. Time-domain based features are the minimum, maximum, mean and variance of the IRA and IRI, the area and length of the respiration signal and the 90th percentile which they define as the value at 90% of the ordered sequence of values inside a signal window. Excluding the percentile, these features were similar to some of those used in this project. The frequency-domain features are all based on the power spectral density calculated using the Fast Fourier Transform, which had previously been proven useful for apneaR detection [1], such as the maximum power in the range of 0.125 - 0.5 Hz. Lastly, the non-linear features encompass measures of complexity or predictability, such as the *Lempel-ziv complexity* or the *Approximate Entropy*. Both metrics describe how “regular” the signal is and thus supposedly help to find the irregular breathing events.

Knorr et al. [29] use a similar set of features (of all three types) to detect general airway obstructions on anaesthetised patients after surgery with the use of a pulse oximeter plethysmograph, which measures changes in the oxygen saturation of the blood. By pressing against the heart during inhalation, the lungs affect the pressure on the heart and therefore the amount of oxygen being pumped through the body at any moment.

Arnold et al. [2] show that the area under the curve of the respiratory signal extracted from this oxygen fluctuation is enough to accurately estimate airway obstruction.

One of the few researchers also working with an accelerometer to detect breathing disorders are Fekr et al. [17]. By using up to two accelerometers calibrated for each patient, they successfully classified eight different breathing patterns such as *Bradypnea*, *Tachypnea* and *Kussmaul*. The authors used similar features to the ones mentioned above, with the addition of the tilt angles of the accelerometer, the tidal volume variability, the phase shift, and a discretised version of the breathing signal called *Symbolic Aggregate Approximation*. The tidal volume variability was previously used by the same authors to design a warning system for acute breathing problems [18]. For the newer publication, they also selected a subset of the features with the *correlation-based feature selection*, although this reduced the performance of most classifiers they tested. The best performance with one accelerometer was reached with *Decision Tree Bagging* with 94.49% accuracy when the sensor is placed on the abdomen, just as in our study.

Sleep apnea and hypopnea seem to be the most popular research objects in the field of breathing disorder detection, probably because of the existing medical equipment (polysomnography) and the clarity of the recorded signal as the subjects are usually sleeping and therefore keeping the disturbances through movements to a minimum. This fact is also the reason why many of the discussed features might prove to be less meaningful for the monitoring during and after exercises, as our subjects are moving around and therefore adding a lot of noise to the signal. None of the above mentioned publications tested their methods on patients outside of laboratory or clinical settings, which makes the applicability of their results for our purpose questionable. The highest classification accuracy of Fekr et al. was obtained on recordings of sitting healthy subjects, who were instructed to breath in a specific way in order to simulate the disorders. As we observed in recordings using the RESpeck, the breathing signals gained from a subject breathing consciously is noticeably more regular and predictable than the ones from subjects subconsciously breathing during other activities.

Nevertheless, many of the features discussed in this section will be adapted and successfully used for our prediction task. The exact features we extracted are explained in Chapter 4.

Chapter 3

Background and preliminary work

In this chapter, we will describe the studies which were the basis for analysis for the experiments in our project. Furthermore, we will go into detail on the data format, as this information will be important for the feature extraction explained in Chapter 4. Lastly, we will provide an introduction to the Hilbert-Huang transform, which is better suited than the Fourier transform for generating frequency spectra from the kind of data we are dealing with, as will be determined empirically in Section 4.2.6.

3.1 Data sources: description of the studies

There are two types of data sources used in this thesis: one is data from real patients, and the other is data recorded specifically for this thesis in order to evaluate methodologies for the feature extraction stage.

3.1.1 Studies on real patients

The Centre of Speckled Computing at the University of Edinburgh has conducted several studies with its Orient and RESpeck sensors (e.g. [16]). Two of those studies were used for our project and will be discussed next.

3.1.1.1 Wester Hailes

The first study was undertaken during pulmonary rehabilitation classes with COPD patients at the *Wester Hailes Healthy Living Centre* in Edinburgh over a period of ten weeks. Pulmonary rehabilitation comprises a series of exercises for the upper and lower body which have been shown to alleviate COPD related symptoms and improve the patient's quality of life [38]. Sixteen patients took part in this study in total and six completed the set of exercises at least four times. During the exercises, the patients wore the RESpeck sensor just below their left rib cage, which sent the recorded movements to an Android application running on a tablet nearby. Information about the subjects is displayed in Table 3.1 (taken from Zhou [49]).

There were seven exercises in total, namely *step-ups, wall push-ups, wall slides, upward row, overhead lift, sit-to-stand* and *walking*. In between the exercises, the patients were asked to rest until their breathing rate went back to their normal level.

We used the data of the six active patients to validate some of the measures like the tidal volume, variance of breathing and breathing recovery rate. This will further be discussed in the corresponding sections in Chapter 4.

Subject	**Age group**	**Sex**	**Years of COPD**
1	70-80	M	17
2	60-70	F	1
3	60-70	F	5-7
4	70-80	M	20
5	50-60	F	<1
6	50-60	F	6

Table 3.1: Information about the participants of the Wester Hailes study.

3.1.1.2 Sutton

This second study is still active and is taking place in Sutton, near London. The patients, aged over 60 years with established COPD condition, were given RESpecks in mid-March 2016 and were instructed to wear the sensor during pulmonary rehabilitation exercises. An Android application (see Figure 3.1) developed by Zhou [49] guides the users through the exercises and makes sure they follow the correct order of exercises,

take breaks in between and are wearing the RESpeck sensor throughout the whole training session. This way, the application stores the exact time intervals of each exercise which makes it possible to extract the periods for data analysis later. As the application can be operated by the patients themselves, they can complete the exercises at home without supervision of a doctor or researcher.

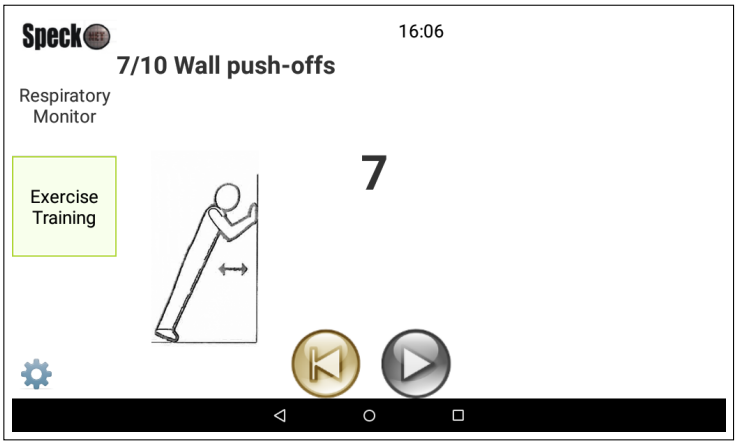

Figure 3.1: Screenshot of the Pulmonary Rehabilitation App used by the subjects to record their exercises.

The exercise session in this study comprised 10 exercises, namely *sit-to-stand, knee extension, squats, heel raises, bicep curl, shoulder press, wall push-offs, leg slide to the side, step-ups* and *walking*. The subjects are instructed by the application to rest in between the exercises and resume the next exercise whenever they feel ready.

In Section 5.2, we will further discuss the statistics about what data we extracted from which patients for the prediction task. It should be noted that this is the first study we know of where COPD patients are monitored unsupervised at their homes. This is accompanied by a lot of uncertainties: does the patient wear the sensor as instructed, at the correct location? Does he follow the exercises as explained by the app? Does he take sufficiently long breaks in between the exercises? All these factors are usually ensured by a researcher, but have to be deduced from the data in this case.

In addition to the exercises, the patients completed a questionnaire which will be further discussed in Section 4.1.

3.1.2 Studies on healthy subjects

Apart from the two studies on real patients, we conducted two studies on healthy patients which were necessary to validate some of the methods for feature extraction.

3.1.2.1 Breath count study

Firstly, we gathered one hour of respiration data from six healthy subjects of ages 19, 20, 22, 23, 49 and 54. The subjects wore the RESpeck sensor below their left rib cage as in the other studies and were recorded in ten intervals of one minute each. Additionally, they counted their breaths for each interval by tapping in a smartphone application. Even though the counting has probably altered the patients breathing pattern compared to subconscious breathing, this was a necessary trade-off, as any external measures would not have returned a "perfect" count. Additionally, as the end time of the recording were specified by the researcher, the last count typically does not denote a full breath, which means that the breathing frequency devised from the breath count will contain a small error due to rounding.

The last recording of subject 4 had to be discarded as the connection to the RESpeck sensor was lost which leads to a dataset of 59 1-minute recordings, each with a corresponding breath count.

3.1.2.2 Coughing study

Secondly, we recorded 87 minutes of breathing data for one healthy subject (age 23) who simulated coughing at irregularly spaced intervals for a duration of 3-7 seconds each. In total, 50 coughing periods were logged with a time tracking application. The coughs varied in frequency and intensity and were performed while sitting (around 95% of the data) and walking (5%).

3.1.2.3 Other recordings

In some sections of this thesis, a RESpeck recording from the author of this thesis is used to demonstrate a specific breathing pattern or algorithm. These recordings were made under idealised conditions and are only meant for demonstration purposes and not

for validation of a specific approach. The details on how they were recorded is therefore irrelevant.

3.2 Data format

The RESpeck stores its recordings into a local database of the smartphone application which is automatically kept synchronised with a server. At a sampling frequency of 12.5 Hz, data points are stored with the following information during pulmonary rehabilitation:

- The timestamp, taken by the phone.
- The magnitude of the acceleration vector in each of the three dimensions x, y and z.
- The breathing signal, as calculated with the approach described by Bates et al. [5].
- The breathing rate, computed a threshold method which will be discussed in detail in Section 4.2.1.
- The activity measure, which is calculated as follows [33]:

  $$act_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}$$

  With x_i being the ith acceleration value of dimension x. In other words, the activity measure captures the difference between consecutive samples as the Eucledian length of those differences.
- The type of exercise being performed at that moment as annotated by the pulmonary rehab application. Default is *no exercise*.

Outside of exercise sessions, only the average respiratory rate per minute is kept, as the unannotated acceleration data would not be of much use. In general, we found that if the approximate movements of the patient, such as during the predefined exercises, is not known, the acceleration data is mostly superfluous. The extraction of the breathing signal only works if the patient is mostly static, with more error introduced the more a person is moving. The RESpeck currently only outputs a breathing rate if the angle change between consecutive acceleration vectors is below a threshold of 0.1 (cf. [5]).

Section 5.2 will go into more detail about how many exercises were recorded by which subjects.

3.3 The Hilbert-Huang-Transform

The HHT [26] is a less well-known technique compared to the Fourier transform for generating a frequency spectrum (power density spectrum) from a time-series signal. The technique is based on the *Empirical Mode Decomposition (EMD)* algorithm which separates a signal into an predefined number of signal bands called *Intrinsic Mode Functions (IMFs)* and one *residue* (or *trend*). The IMFs and the residue sum back exactly to the original signal. The IMFs each capture different frequency components and in contrast to the Fourier transform, two bands can contain the same frequency values at different times (although not at the same time). Figure 3.2 shows an example of an EMD on a breathing signal. As can be seen from the figure, the IMFs lower in the figure contain the lower frequency components, whereas the top ones contain the higher frequency components. Every signal can be decomposed into an arbitrary number of IMFs, although on our breathing data, after about 5-6 IMFs, further IMFs are close to zero.

As the IMFs do not focus on a single frequency component, but can span a range of them, they are ideal for extracting sub-components of the signal with physical meaning. The EMD has been successfully used on a pressure sensor drawn over a subject's mattress to extract both the Electrocardiogram (heart signal), as well as the respiratory signal as separate IMFs [9].

Due to the construction of the EMD algorithm, the IMFs fulfil certain properties: "(1) in the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one; and (2) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero." [26]. These properties allow the definition of an *instantaneous frequency*, i.e. a frequency value local in time. This local frequency is calculated by the derivative of the analytical component of the IMF which is computed through the *Hilbert transform*.

Taking the instantaneous frequency signals from all IMFs and combining them with the amplitude values of the signals into a three dimensional plot creates a frequency spectrum called the *Hilbert spectrum*. Integrating the Hilbert spectrum by time into a

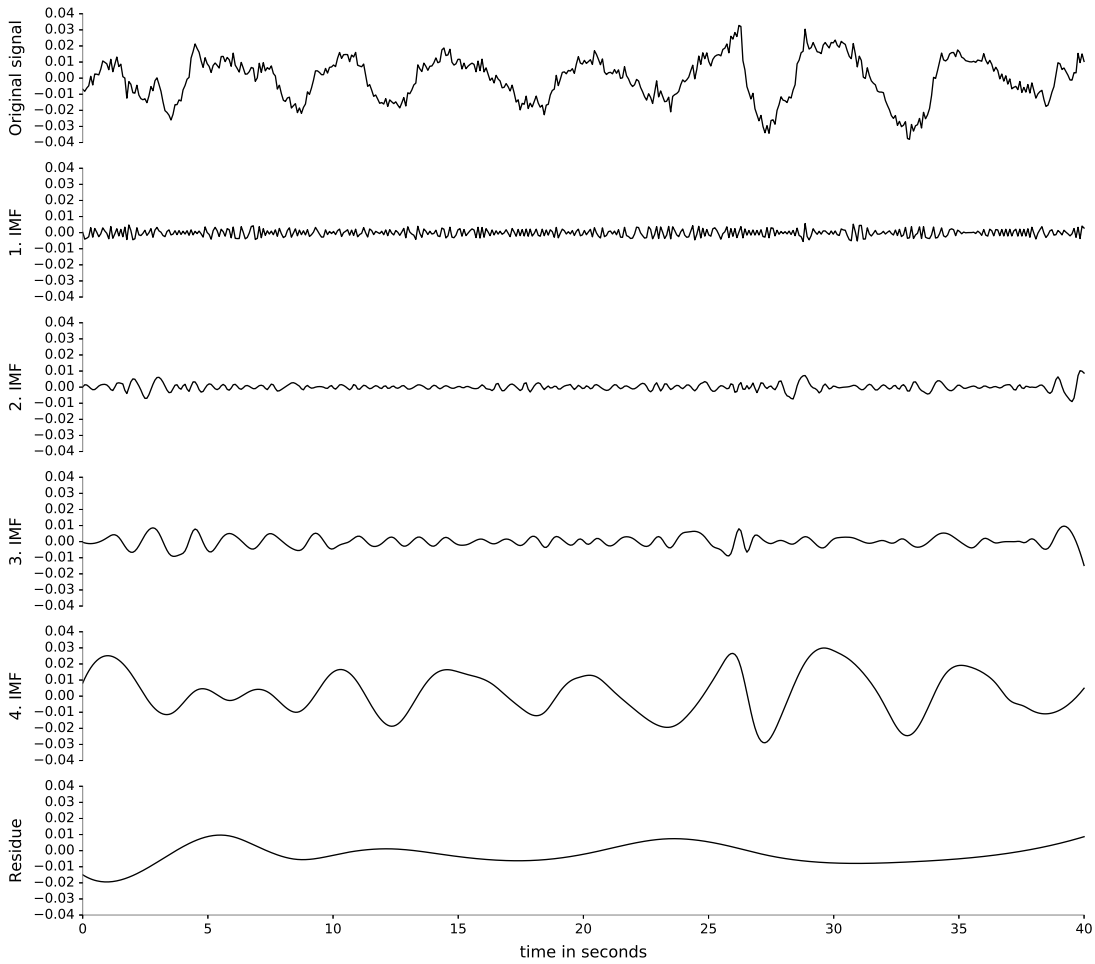

Figure 3.2: A breathing signal (top) with its Intrinsic Mode Functions and the Residue

two-dimensional plot then leads to the *marginal Hilbert spectrum (MHS)*. This MHS is very similar to the Fourier spectrum (FS) in that it displays the prevalence of frequency components in the original signal. Huang et al. state that the residue can be either included or excluded from the MHS calculation depending on the application. We found the results to be superior on our data when including it. Figure 3.3 shows the Hilbert spectrum, as well as the MHS in comparison to the FS for the same breathing signal decomposed in Figure 3.2.

As already mentioned, the EMD is an algorithm and therefore not as well mathematically understood as the Fourier transform, which means that its effectiveness could only been proven empirically so far. Through its construction, however, it has the big advantage of being able to analyse non-linear and non-stationary signals which are the norm in real-life sensor data, in contrast to the Fourier transform which only works on linear and stationary signals. Examples can be found in Huang et al. [26].

Salisbury and Sun [39] used the Hilbert spectrum to discriminate between patients with and without apnea. They determined a *critical frequency* of 1.5 Hz which acts like a threshold. Spectra with frequency components noticeably above this threshold were classified as containing apnea.

For this thesis, the MHS was used to estimate the speed of execution of the exercises. Details can be found in Section 4.2.6.

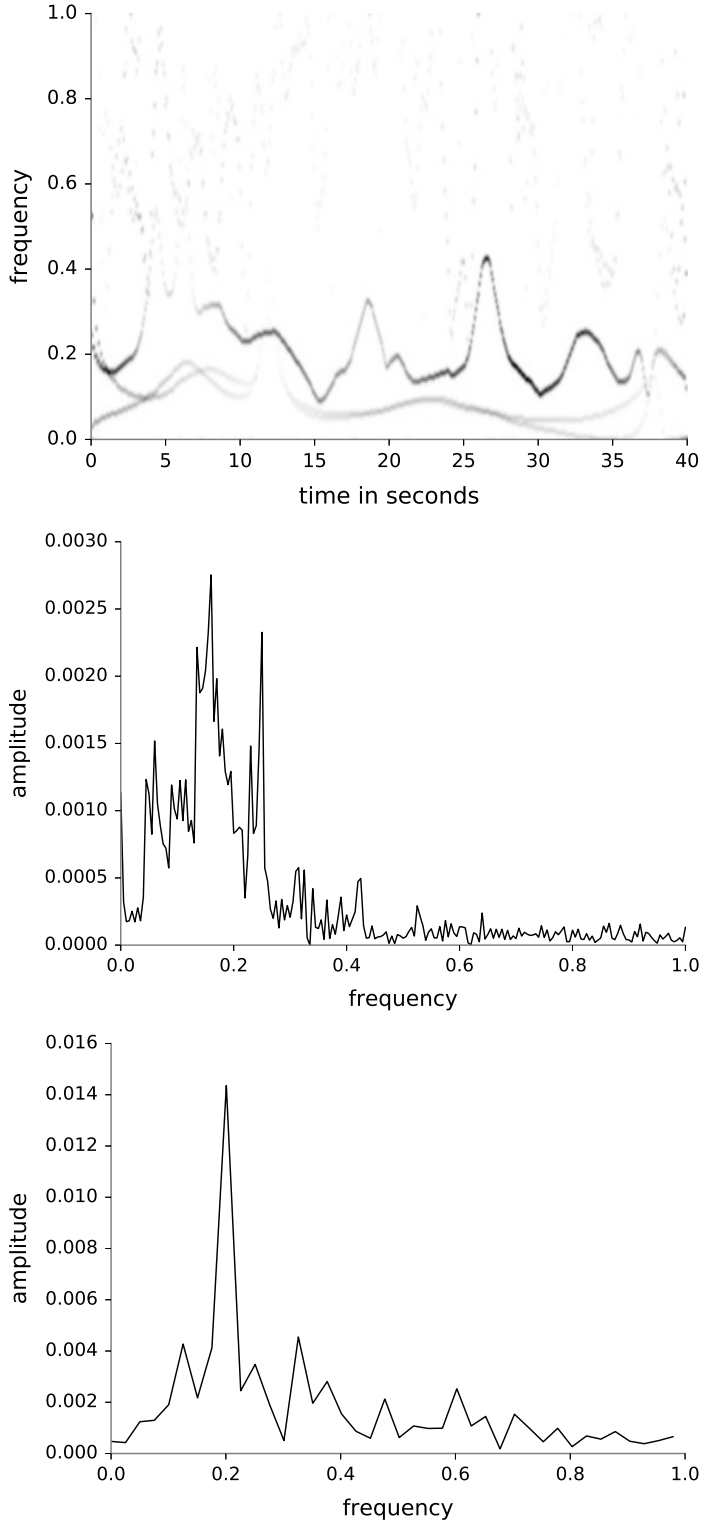

Figure 3.3: The Hilbert spectrum (top), the marginal Hilbert spectrum (middle), and the Fourier spectrum (bottom) for the breathing signal in Figure 3.2

Chapter 4

Labels and feature extraction

This chapter will examine the kind of features extracted from the breathing signal, with justification for selecting them and the techniques for computing them from the raw sensor data. These features will be used later to predict the CAT questionnaire which was introduced in Chapter 2, and which will now be described in detail here.

4.1 The COPD Assessment Test

The CAT is a questionnaire comprised of eight questions which were selected based on their ability to capture the patient's state of health. Interviews with physicians and pulmonologists led to an initial set of 21 questions which were than ranked based on different statistical measures, such as how well they were understood by the patients, how well the question correlated to the assessments made by physicians, how strongly they correlated with other questions in the set and whether the scores given by patients were evenly distributed [27]. The resulting eight questions are all based on a six-point scale from 0 to 5, where 0 represents the best possible answer and 5 the worst, and are posed so that the question itself is implicit through the labels of the extremes:

1. I never cough $\cdots$ I cough all the time
2. I have no phlegm (mucus) in my chest at all $\cdots$ My chest is completely full of phlegm (mucus)
3. My chest does not feel tight at all $\cdots$ My chest feels very tight

4. When I walk up a hill or one flight of stairs I am not breathless ··· When I walk up a hill or one flight of stairs I am very breathless
5. I am not limited doing any activities at home ··· I am very limited doing activities at home
6. I am confident leaving my home despite my lung condition ··· I am not at all confident leaving my home because of my lung condition
7. I sleep soundly ··· I dont sleep soundly because of my lung condition
8. I have lots of energy ··· I have no energy at all

This questionnaire was implemented in a mobile application which the patients are instructed to complete daily. After completion of the questionnaire, the answers to every score are stored, together with a timestamp, in a local database synchronised with the central server. In contrast to the mobile application deployed by Smith et al. [42], the questionnaire in this study was not intended as an early warning system, but rather as a means to provide training labels for the features extracted from the RESpeck sensor data. The goal is to predict the CAT score without any intervention by the patients apart from doing the pulmonary rehabilitation exercises. Based on patient feedback, our initial implementation of the questionnaire was slightly modified by the study administrators to make the answering mechanism more obvious.

4.2 Selecting and extracting features

In order to predict the CAT scores, features need to be extracted from the RESpeck signal. The following sections will look at each of these features in detail.

4.2.1 Respiratory rate

4.2.1.1 Generating the breathing signal

The first step to extracting any time-domain feature is to reduce the three-dimensional accelerometer signal into a one-dimensional breathing signal. The resulting reduction process should throw away as much noise as possible, i.e. acceleration data not related to breathing, while ideally only keeping the breathing signal. Of course, the distinction

between the breathing motion and the movement of other body parts is not clearly set, making it impossible to measure exactly the exhaled and inhaled volume as one would using a spirometer. However, one can try to minimise the difference between the reduced signal and the actual breathing signal. A few factors have to be considered for such a method:

1. No single dimension contains the whole breathing motion. Even though most of the breathing will occur in the z-dimension as it is the one perpendicular to the abdomen, information would be lost by only considering this dimension. Breathing motion does not follow a straight line and depending on the position of the subject and his breathing style, the chest or the abdomen will play a bigger role in the motion which means the sensor will tilt instead of just shifting on a straight line.

2. The direction of the breathing can change, i.e. a peak can sometimes mean a maximum inhale and sometimes an exhale [16].

3. The acceleration signal can show breathing movement when the airways are physically obstructed, even though no actual breathing takes place [6].

4. Accelerations due to breathing are tiny and usually overshadowed by other movements of the body. This means that the more static a person is during the recording, the more accurate it will be. For this reason, the RESpeck is intended to only measure the quiet breathing at rest [5].

The first point implies that all three axis of the accelerometer should be used to calculate the combined signal. The second and third points mean that the resulting breathing signal cannot simply be judged visually, but has to be evaluated with an external measure. Bates et al. [16] validated the original Orient sensor by comparing its signal to that of a nasal cannula, although a good match with the cannula signal does not necessarily enable an accurate calculation of the respiratory rate, as will be shown later. In this thesis, the signal is assessed by counting the breaths taken during each recording. The focus here was on getting an accurate respiratory rate, although this probably also leads to a reasonably accurate signal shape . Lastly, there has to be some form of test as to whether the subject is moving above a certain threshold level which would render the signal unacceptable.

The following paragraphs will look into the current and alternative methods for generating the breathing signal. For each method, strengths and weaknesses are discussed and lastly, the signal best suited for the respiratory rate calculation is evaluated.

4.2.1.1.1 Angular velocity For the original Orient sensor, Bates et al. [6] devised a technique to generate the breathing signal based on the change of rotation of the sensor. The idea behind this approach is that the linear motion of the sensor on the abdomen is small compared to the angle change of the gravitation vector. As the angle changes are unpredictable during static periods, the average angle $\vec{a}$ and average rotation axis $\vec{r}$ in a sliding window are calculated. For each timestep, the angle between the current vector and $\vec{a} \times \vec{r}$ is computed. Lastly, the derivative over time of these angles calculates the angle changes over time, or *angular velocity*.

Although the resulting signal has been shown to correlate closely with a nasal cannula signal, the method has two drawbacks: Firstly, this method assumes that the breathing can be solely extracted from the angle changes without taking the linear changes along a specific dimension into account. As we will in Section 4.2.1.3, the linear changes alone are a very good approximation of the breathing signal and actually lead to a better breath count. Secondly, by using the derivative of the angle changes, the number of zero-crossings can be higher than in the actual breathing. Also, the phase of the signal will shift slightly, which is not important in our use case. Lastly, due to the construction of the signal through comparison with a mean rotation angle and axis in fixed windows, the first and last $\textit{window size}/2$ points are lost during the extraction and are therefore set to zero. This can be circumvented by only using part of the window for the edge cases, as will be done for another signal (Section 4.2.1.1.3). Figure 4.1 shows the angular velocity in comparison to the other approaches mentioned in the following on a breathing signal for which the subject took breaths of decreasing volume. The higher number of zero crossings and the cut off edges are clearly visible in this example.

4.2.1.1.2 The derivative of PCA As the task at hand is one of dimensionality reduction from the three-dimensional acceleration signal to a one-dimensional breathing signal, it seems plausible to use *Principal Component Analysis (PCA)* to perform the reduction. PCA is a well-known technique to reduce dimensions based on maximising the resulting variance of the samples in the reduced dimensions [8]. The larger the variance, the more discriminative information is kept after the reduction process. The

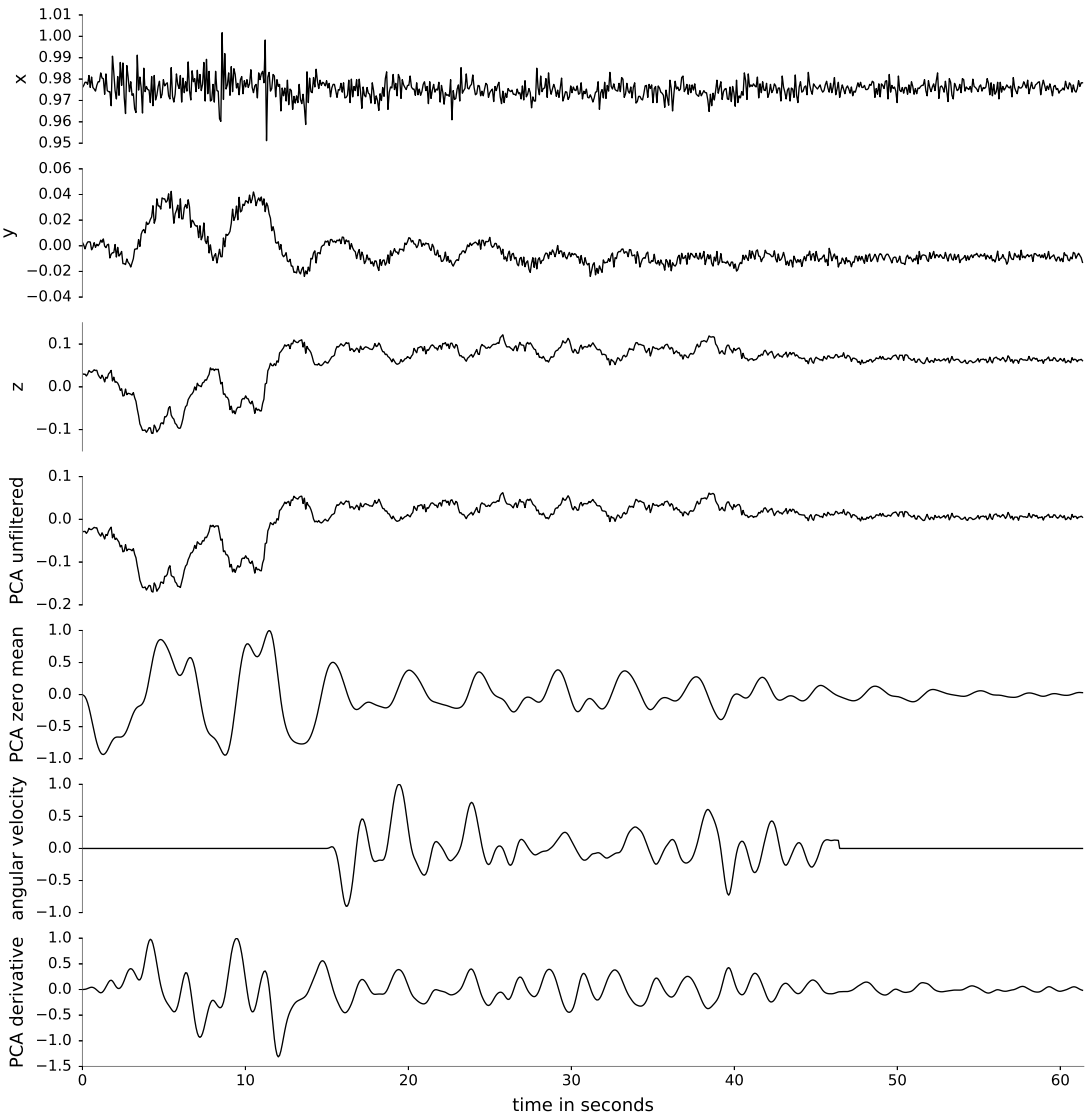

Figure 4.1: The accelerometer signal of a 62 second recording, the PCA reduced signal, and three different breathing signals extracted from it below

intuition for applying it on the accelerometer signals is that when a subject is sitting or standing still, the breathing direction will probably be the dimension of maximum variance. PCA will therefore maximise the amount of linear breathing kept while throwing away the noise in other dimensions. Note that this reasoning contradicts the one made originally for the Orient sensor, where the linear movement was assumed to be too small to be informative.

Figure 4.1 also displays the unfiltered PCA reduced signal. At first sight, it seems almost identical with the z-axis of the accelerometer. Looking at the span of values in each of the accelerometer dimensions, we can confirm that the z-axis indeed spans a

range of values much bigger compared to the other two dimensions. The PCA signal is not a perfect match, however, which implies that the breathing axis lies somewhere close to, but not on the z-axis.

The PCA signal itself cannot be used as a breathing signal, however, as it is not centred around 0. One way to accomplish this is to take the derivative of the signal, just as in the case with the angular velocity. The result is termed *PCA derivative* in Figure 4.1. In contrast to the angular velocity, it now spans the whole time range. It still has the higher number of zero crossings, however, which are even more distinct than with the angular velocity and clearly too many when compared visually with the PCA signal.

4.2.1.1.3 PCA with a zeroed mean The third approach tries to mitigate this problem by centring the PCA signal around the axis $y = 0$. The intuition is that between seconds 20 and 40 in the PCA signal, the graph looks like a "normal" breathing signal, albeit flipped (explanation below). We should therefore be able to subtract a constant from this part of the signal to get it into the right shape where this "constant" is most likely the mean of the signal inside a window.

In order to transfer this approach to the whole signal, a running mean in a fixed window size is subtracted from the PCA signal, i.e. for every point p with index i in the signal, the mean of the interval $\left[max(i - window\_size/2, 0), min(i + window\_size/2, length(signal))\right]$ is subtracted from p. Therefore, at the signal edges, only part of the window is used to calculate the mean.

The final result was flipped as that got better results in the evaluation, which implies that the flipped version corresponds better to the actual breathing, i.e. peaks implies inhalation, valleys exhalation. Additionally, the subject in this recording made longer pauses during the exhaled state which means that the peaks in the signal should be shorter, as is the case in the flipped version.

Lastly, the signal is filtered with a fourth-order Butterworth bandpass filter (highcut of 0.5 Hz) which introduces a slight phase shift.

In Figure 4.1, the PCA zero mean signal seems to have the best correspondence with the acceleration signal. Judging the signal visually is not enough, however. In the following, several methods for calculating the breathing rate from the signal are introduced. The different signal generation techniques are then evaluated together with the respiratory rate calculation methods on a study conducted specifically for this purpose.

4.2.1.2 Calculating the respiratory rate

Next, the respiratory rate can be computed in the signal generated. There are three main approaches to achieve this:

1. Count the breaths in a period and divide this number by the length of the period in minutes to get the breaths per minute (bpms) value.
2. Determine the length of individual breaths in seconds, extrapolate each breath to the *breaths per minute* value by dividing 60 seconds by the breath length, and lastly take the mean over all extrapolated bpms.
3. Use frequency-domain spectra to determine the mean power or maximum power in a frequency range.

For the first two methods, the question remains of how to find the breath count or the length of individual breaths. There are multiple ways to achieve this, two of which will be explained and evaluated here.

4.2.1.2.1 Peak find algorithm The python package *PeakUtils* [37] determines the peaks of a signal based on a *threshold* and a *minimum distance* parameter. The threshold value determines how high a peak must be compared to the maximum value in a signal in order to be considered a peak. All local maxima above the threshold level are taken into account and sorted in decreasing order by the amplitude value. They are then looked at one-by-one. If there is no other peak inside the minimum distance spanned around the currently examined peak, the current peak is kept, otherwise discarded.

With suitable parameter values, the results for this algorithm are similar to what a human might label as a “global peak”. The peaks themselves can be used for counting, while the difference between peaks would be a measure of the breath duration.

One problem of this approach is that the two parameter values are static for the whole period being analysed, as well as for different subjects. This means one would have to find universally applicable parameter values which seems infeasible considering the completely different styles of breathing across patients.

4.2.1.2.2 Root-mean-square threshold Instead of these predetermined threshold values, a dynamic threshold value based on the local properties of the signal would be

preferable. Bates et al. [5] used the root-mean-square (RMS) inside a predetermined window as a positive and negative threshold value which adapts to the specific form of the signal. The window size has to be predetermined and was found to be optimal at 42 samples (see Section 4.2.1.3). Figure 4.2 displays the same breathing signal as in Figure 3.2 with the RMS threshold (window size: 42) and the peaks from PeakUtils (peak threshold: 0.5, minimum distance: 25). As can be seen in the Figure, the RMS adapts to the decreasing amplitude size.

After having calculated the RMS thresholds, the breaths can be extracted by iterating over the sample points and checking whether they are above or below the threshold values. A full breath is counted after the signal has surpassed the top threshold, then fell below the lower threshold and lastly crossed the top threshold again. The two vertical black lines in the figure denote the first breath extracted this way. Just as described above, the number of breaths or the breath length can be used to calculate the bpms.

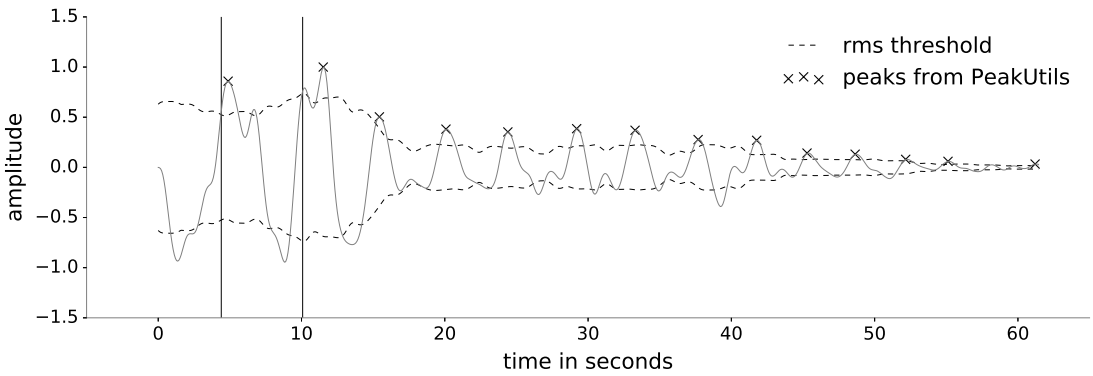

Figure 4.2: A breathing signal with the thresholds determined by the RMS and the peaks from PeakUtils. The black lines denote the beginning and end of the first breath extracted with the RMS threshold.

4.2.1.2.3 Frequency-domain based methods Instead of looking at local properties in the signal, the Fourier spectrum (FS) or the marginal Hilbert spectrum (MHS) (see Section 3.3) can be analysed. Koley et al. [30] determined the respiratory rate by taking the maximum power in the frequency range of 0.125 - 0.5 Hz. However, they did not seem to validate this calculation empirically. We found that the maximum power is inferior to the above mentioned time-domain techniques by a wide margin. This can be explained by a simple example: the breath durations typically vary quite substantially, even for the same patient and in close time proximity. Assuming we have a patient who breathes half of the time at 0.24 Hz, and otherwise with 0.28 Hz, we would say the

mean bpms is 0.26. The maximum power, however, will either lie at 0.24 or 0.87 Hz which is a difference of 0.02 Hz or $0.02 * 60\ sec = 1.2$ bpms.

Taking the mean of the power values between 0.24 and 0.28 Hz would possibly solve the problem for this case. However, it is impossible to generalise this method, as we cannot specify a frequency range which will lead to good results for all patients and recordings. For the same frequency spectrum even a small change in the frequency range will change the resulting mean bpms substantially.

Figure 4.3 displays the FS and MHS for the same breathing signal as in Figure 4.2, together with a vertical line marking the actual frequency as counted by the subject. One of the MHS power spikes is very close to this actual value and the deviation could simply be due to the lack of precision with counting full breaths manually (compare Section 3.1.2.1). However, it is not the maximum power spike and would therefore have not been taken into account by the method due to Koley et al. [30].The FS similarly has a spike near the actual frequency, but not the biggest one.

4.2.1.3 Validation of the respiratory rate calculation

In order to evaluate the breathing signal and respiratory rate calculation, a study with six healthy patients was conducted, see Section 3.1.2.1 for details.

The breath count of the subjects enables the calculation of target values for the respiratory rate: the duration of the period in minutes is the number of samples in that period, divided by the sampling frequency (to get seconds), divided by 60 (to get minutes). The respiratory rate in bpms is the breath count divided by the duration in minutes.

These target values were compared to the output of the above-mentioned techniques: Angular velocity, PCA derivative and PCA zero mean, each with PeakUtils and threshold-based peak detection, as well as the MHS- and FS-based techniques on each of the above signals. The angular velocity signal was not used for the frequency spectra and the peak count because of the missing edge values. The error measure was the mean absolute error/deviation (MAE) of the predicted and target value in all the periods.

As expected, MHS and FS each led to unsatisfactory MAE scores of 7.2 and 11.4, respectively on the PCA zero mean signal, and 7.72 and 12.1 on the PCA derivative signal, both on 5 IMFs and a frequency range of 0.05 to 0.5 Hz. They returned similar results for different numbers of IMFs and frequency ranges. We could not adopt the

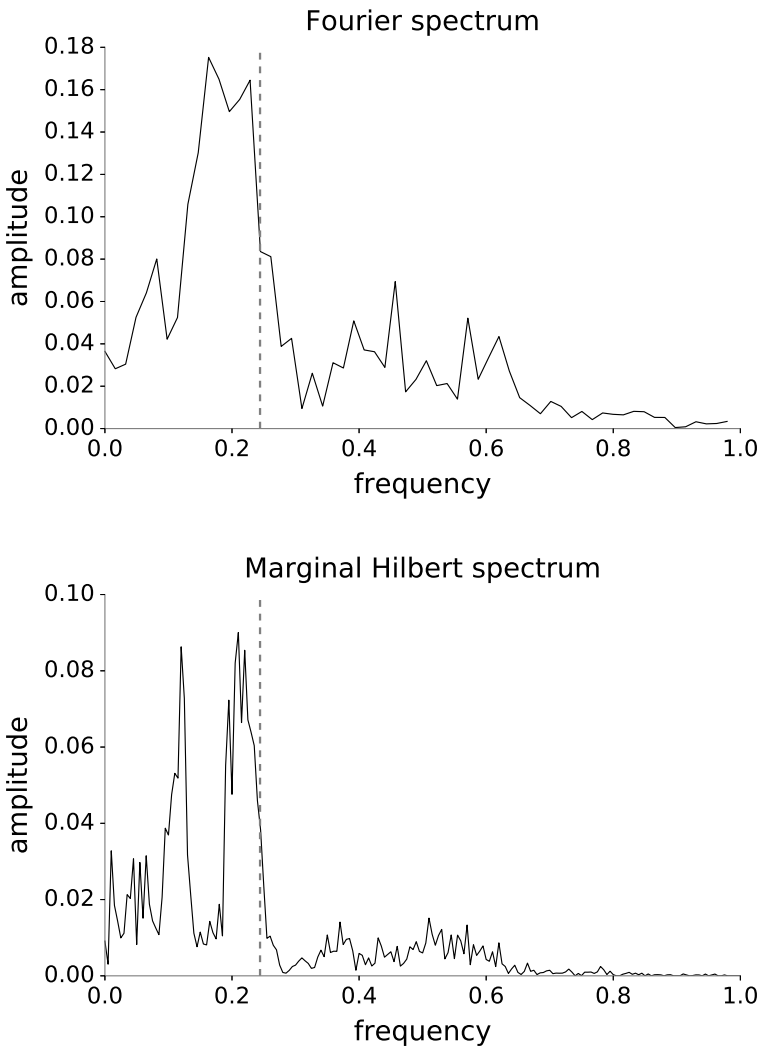

Figure 4.3: The Fourier spectrum and the marginal Hilbert spectrum for a breathing signal including the actual breathing rate as measured by the subject, displayed as a dashed line.

frequency range from Koley et al. as one of our subjects had a respiratory rate of only around 0.1 which would have fallen below their minimum frequency – another example for the impracticality of having to define a frequency range for all subjects.

For the time-domain techniques, grid search was used to find the best parameters. As the subjects were all healthy, the respiratory rate was relatively low throughout the whole study. In the actual analysis, the subjects will have completed exercises before each period being analysed which means that their respiratory rate will be elevated. Therefore, the minimum distance parameter for the peak count which depends on the typical distance between peaks should be interpreted with care. Additionally, the

optimal highcut value for the Butterworth bandpass filter was found to be at 0.15 Hz which is obviously too low for elevated breathing. It is difficult to find a reliable source on the typical breathing rate after exercises, so for this thesis we assumed a maximum rate of 40 bpms based on own breath count in rest periods after high intensity exercises. This corresponds to a highcut of 0.66, which was rounded to 0.65.

With the highcut set, the optimal parameters for PeakUtils were 0.36 for the peak threshold, 47 for the minimum distance between peaks, 5 for the filter order, and 65 for the window size of the running mean in the PCA zero mean calculation. This resulted in a MAE of 1.56, using breath extrapolation to get the bpms value.

For the threshold technique, the best parameters were 30 for the RMS window size, 63 for the running mean window size and 4 for the filter order. This resulted in a MAE of 1.11 on the PCA zero mean signal, with the count technique to determine the bpms value. The second best option with a MAE of 1.19 was also on the PCA zero mean signal, but with the extrapolation technique and slightly changed parameter values, namely 40 for the RMS window size and 69 for the running mean window size.

In all cases mentioned above, the PCA zero mean signal was flipped, i.e. multiplied by -1 before the calculation. It is therefore assumed that the breathing direction in the flipped version is such that the peaks denote inhalation and the valleys exhalation events. In the following, when we speak of the PCA zero mean signal, we will therefore always refer to the flipped version.

Lastly, the method currently used which is a simplified version of the angular velocity calculation resulted in a MAE of 2.65, i.e. considerably worse than the results above. Additionally, the current method does not generate a continuous breathing signal which made the implementation of a new technique for this project unavoidable.

The results are summarised in Table 4.1. From these results we can conclude that the best approximation of the actual respiratory signal is the PCA zero mean signal. Additionally, the best technique to extract the respiratory rate from the signal is using the total count of the breaths as detected with the RMS threshold.

4.2.1.4 The respiratory rate as a feature

The respiratory rate and its change over time is assumed to be a very good indicator of a patient's health (cf. [5, 19]). Due to its assumed importance, it was recorded as a feature

Domain	Peak detection	Bpms calculation	MAE
Time-domain based	PeakUtils	Count	1.65
		Extrapolation	1.56
	Threshold	Count	1.11
		Extrapolation	1.19
Frequency-domain based	MHS	Max power 0.05-0.5 Hz	7.2
	FS	Max power 0.05-0.5 Hz	11.4

Table 4.1: Results for different respiratory rate calculation techniques.

for each rest period separately, in contrast to all other rest features which were only extracted as mean or variance over all the rest periods. This means that for some rest periods, which were too short to even extract one breath, the value had to be guessed. The guess was chosen to be the mean of all other rest periods in that exercise block. Even though this will distort the results, the benefit of the increased level of detail in the respiratory rates will likely outweigh any induced error.

In addition to the mean respiratory rate for each rest period, the total mean of all rest periods in the exercise block, as well as the mean of the variance of the bpms in each period was recorded as a feature. In contrast to the mean respiratory rates which used the total count of breaths and the parameters as denoted above for the lowest MAE score, the variance was computed on the best parameter values for the extrapolated bpms. The reasoning behind this is that the optimised parameters for the extrapolated bpms resulted in the best approximation of individual breaths rather than the combined breaths of a period which makes the variance value more reliable.

In total, this lead to a count of 12 features related to the respiratory rate.

4.2.2 Recovery after exercises

We could not find any scientific literature on breathing rate recovery after exercises and its significance for health, which probably has to do with the lack of suitable devices which can be worn during exercise periods. However, research shows that the recovery of heart rate after exercises is a good predictor for the mortality of a patient [12]. This suggests that the changes in breathing after exercises should also be an indicator of the well-being of the patient.

There are two ways to measure the rate of recovery after an exercise: by the change of the respiratory rate over time in a rest period, and by the total length of the rest period.

4.2.2.1 Respiratory rate recovery slopes

For this work we approximated the recovery of the breathing rate through linear regression. Although the actual recovery curve will probably follow a more complicated polynomial pattern, linear regression makes more sense here. Firstly, polynomial curves are likely to overfit with an unpredictable behaviour at the edges of the recording. Secondly, the parameters of polynomial functions are a lot more difficult to interpret than a simple slope of the linear regression line. Thirdly, the respiratory rates themselves do not match the real values perfectly and any fit more complicated than a linear one will be more severely affected by these errors.

The input parameters to the linear regression fit are the extrapolated bpms value for each breath in a rest period as y-values, and the centre of those breaths as x-values, which were defined as the index of onset of the current breath, plus the index of onset of the next breath, divided by two. It is assumed in our study that the subjects do not rest for a prolonged time after their breathing rate has recovered, as this would flatten the slope and therefore distort the value.

Figure 4.4 shows an example of a breathing signal from the author of this thesis after high intensity exercises. The recovery of the respiratory rate as measured by the PCA mean zero signal follows a clear linear downward trend.

In order to justify the inclusion of the recovery slope as a feature, the method was tested on a previous study in Wester Hailes, Edinburgh, which was introduced in Section 3.1.1.1. In this study, six patients completed pulmonary rehabilitation exercises just as in the main study of this project. Calculating the recovery slope on the rest periods of each patient leads to the graphs depicted in Figure 4.5. As can be seen from Table 3.1, patient 1 has the longest history of COPD. However, this patient surprisingly also has some of the steepest recovery slopes. Even though the mean of the slopes of patient 1 and 3-5 are negative, the figure does not seem to match the actual severity of the patient's disease very well. This could have several reasons: firstly, the recovery slopes could simply not be a good predictor of the patient's health. Secondly, the breathing signal from the RESpeck might contain too much noise to accurately extract the slopes.

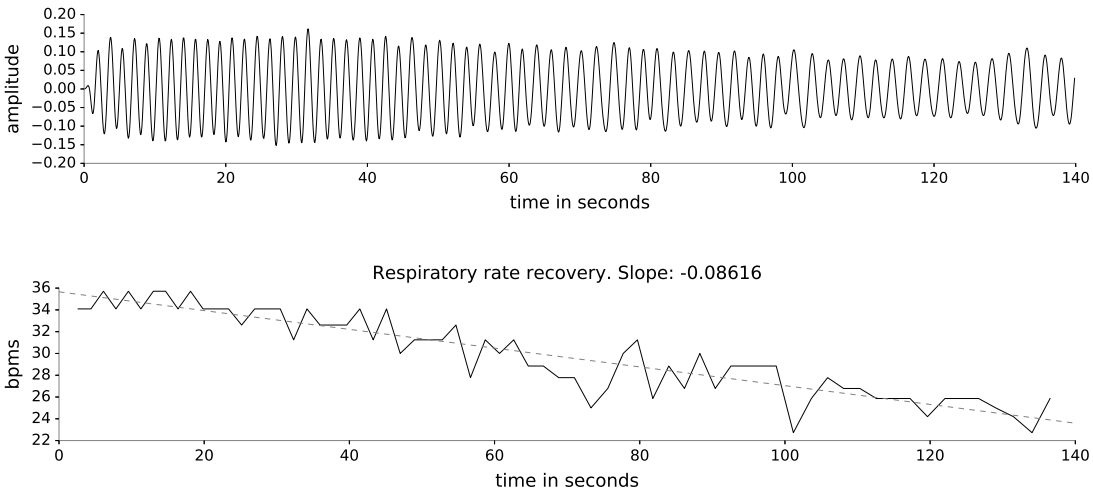

Figure 4.4: An example of a breathing signal after high intensity exercises with the respiratory rates and the linear regression line below.

Lastly, the data basis in this study could be too small to get meaningful results, as it only contains 2-6 exercise blocks for each patient.

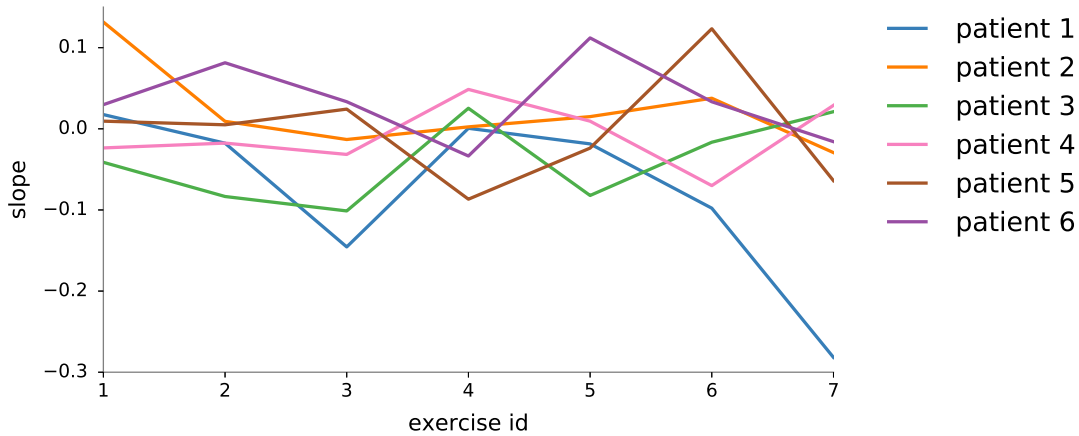

Figure 4.5: Mean recovery slopes in rest periods after exercises for each patient.

Even though the recovery slopes do not seem to have a predictive power in the Wester Hailes study, we decided to include them as feature nonetheless, as the comparable heart rate recovery has shown to be an effective predictor and the results of the Wester Hailes study might not be significant due to the reasons mentioned. Only the mean slope over all rest periods was included in order to compensate for irregularities in individual periods.

4.2.2.2 Length of rest periods

The length of a rest period is determined by the patients themselves who are instructed by the Android application to rest until they feel capable to continue with the exercises. This way, the length is a subjective measure of the speed of recovery. The mean length of all rest periods as the number of samples was therefore included as feature.

4.2.3 Tidal volume

The term *tidal volume* describes the amount of air which is inhaled or exhaled during normal breathing. As a typical sign of COPD patients is shallow breathing [21], the mean of the tidal volumes in a rest period could provide information about the obstruction of the airways. Additionally, the statement "My chest does not feel tight at all" from the CAT possibly correlates well with the shallowness of the breathing. Apart from the mean value, an increase in the variability of the tidal volume has been found to be a better indicator of opioid-induced respiratory depression in children than a decreasing breathing rate [4]. Fekr et al. [18] devised an alarm system for detecting five different breathing disorders based on the tidal volume variability calculated from an accelerometer signal. This volume measurement correlated by 0.87 with the volume changes extracted in parallel from a spirometer. The best accuracy achieved was 98.28%, with high true positive and true negative rates. As the data basis and the goal of the study was similar to our work, the approach of Fekr et al. was adopted for this project, on the assumption that a the tidal volume variability might also allow deductions about the general state of a COPD patient. The main ideas of this method are explained next.

The tidal volume in an accelerometer signal is defined as the difference between a peak and the mean amplitude of the two adjacent valley points:

$$TV_i = p_i - \frac{v_i + v_{i+1}}{2}$$

Where p_i and v_i are the value of the ith peak and valley point in the signal, respectively as illustrated in Figure 4.6. This formula would likewise be used for a spirometer signal.

The peaks and valleys are computed by first extracting the individual breaths with the threshold method and parameters of the extrapolation technique discussed in Section 4.2.1.3. The peak is then set as the maximum value of the first part of the breath which is above the threshold for the reason that simply taking the maximum value over the

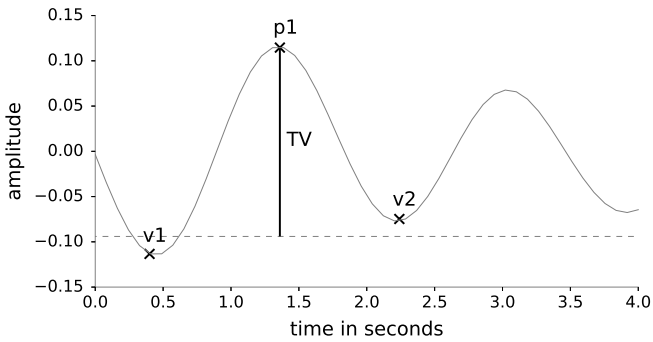

Figure 4.6: Calculation of the tidal volume (vertical black line) with the peak and mean of adjacent valley points (dashed line).

whole breath could select the endpoint of the breath which is not a real peak. The valleys are set as the minimum value in a breath. After all the volumes in a period are calculated this way, we store their mean and variance. The mean of all the mean and variance values for all periods are then added as one feature each.

Fekr et al. propose a slightly more complex variant of computing volume variability by taking the slope of the tidal volume values in windows of three volume samples. In this way, a single outlier is compensated for by the other two values in the window. They continue with these slope values by binning them into discrete value ranges based on the total distribution of the slopes. They then define an alarming state as one in which the slope enters the edge bins, i.e. very unlikely ranges [18]. As the goal of this project was not to issue discrete alarms but rather to detect continuous state changes in the patient's health, we modified this approach by including the total mean of the variance of the volume slopes for each period. This resembles the simple variance of the volumes closely but might be less sensitive to outliers.

In total, we have three features related to the tidal volume: the total mean of the volume mean, the volume variance, and the volume slope variance for each period.

4.2.4 Activity features

The activity level of patients with COPD as measured by their time spent outside has been shown to be a good (negative) predictor for exacerbations [15]. Although there to our knowledge no research on the topic of activity level in rest periods after exercises, we hypothesised that the less exhausting an exercise feels for a patient, the more he will move around during the breaks. Similarly, we suspected that the healthier a patient,

the more active he would be during exercises themselves. Therefore, three measures capturing the activity level of the patient were extracted for both rest and exercise periods:

- In Section 3.2, the **activity measure** defined by Bates et al. [33] was introduced. As a reminder, the activity level there is calculated by the length of the vector difference between two consecutive vectors in the accelerometer data:

$$act_i = \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}$$

- In the process of calculating the angular velocity, the **angle change** between two consecutive vectors is used to filter out periods where the subject is moving too much. This angle change is defined as follows:

$$\theta_t = \cos^{-1}(a_t * a_{t-1})$$

  where a_t is the acceleration vector at time t.

- If the subject is not static, the breathing signal will typically be concealed by other movements, meaning that the absolute amplitude values as used by the tidal volume calculations will not have any interpretation related to breathing. It is not known whether the subjects of our study are static during breaks between the exercises or keep moving around, although they are advised to sit down during the rest periods. In case they do move around, the amplitude values are better interpreted as measures of the activity level. Additionally, it makes more sense to extract this activity measure from the unfiltered PCA signal and use the **variance of all amplitude values** instead of only the peaks.

All three of these activity measures are extracted from both the rest as well as the exercise periods, by taking the mean of each feature over all periods in a block.

4.2.5 Coughing

Coughing is one of the few local patterns which are very likely a good indicator for the health of the patient, in particular as the amount of coughing is a direct question in the CAT itself ("I never cough $\cdots$ I cough all the time"). Intuitively, coughing should be clearly distinguishable from the normal breathing as long as the patient is relatively static and the accelerometer sensor is not tapped on during any movements.

In order to devise a model for detecting coughing patterns, a recording of a single healthy subject was made over a period of 87 minutes in which the subject coughed in 50 irregular time intervals for a length of 3-8 seconds and recorded the time of the coughs. Figure 4.7 shows three 8-second periods from this recording, each represented in three graphs with the unfiltered PCA signal on the top, the derivative of the z-dimension in the middle, and the angles between consecutive vectors at the bottom. Graphs displaying the same type of information are depicted with the same value range to make the graphs visually comparable. The first two periods were recorded while the subject was sitting, while the last one was taken from a time span when the subject was walking.

The detection of coughs has to be made on unfiltered data as a filter would smooth the spikes which are characteristic of the coughs. The z-dimension of the accelerometer was chosen as it is the one pointing away perpendicular from the subject's abdomen for all subjects wearing the RESpeck. The outburst of air during a cough is generated by a sudden, fast contraction of the diaphragm which is just below the usual location of the RESpeck. The cough should therefore mostly follow the direction of the z-dimension, just as this dimension also contains most of the breathing information (see Section 4.2.1). Focusing on this dimension also possibly reduces the noise from sudden movements into other dimensions which are not related to coughing, such as the impacts with the ground during walking.

Illustrating the coughing period in Figure 4.7, we see that the sampling frequency of 12.5 Hz is too low to detect any distinct pattern apart from simple spikes and the height of those spikes – an observation which is confirmed by other coughing periods. Additionally, spikes generated at the end of the movement period look very similar to the ones in the coughing period. The question therefore arises whether it is possible to even distinguish coughing from other movements such as walking at this sampling rate.

In order to answer this question, two techniques were devised and tested on three different signals. The signals are the unfiltered PCA of the acceleration signal, the single z-dimension, and the angles, as defined in Section 4.2.4. As the absolute amplitude value will vary with patients and by how much the patients are moving at the time of the cough, we take the derivative of the PCA and z-signal. Small value changes at any amplitude level will get a low value in the derived signal, whereas the spikes will remain as high spikes (cf. Figure 4.7), which is just what is needed here. The angle calculation

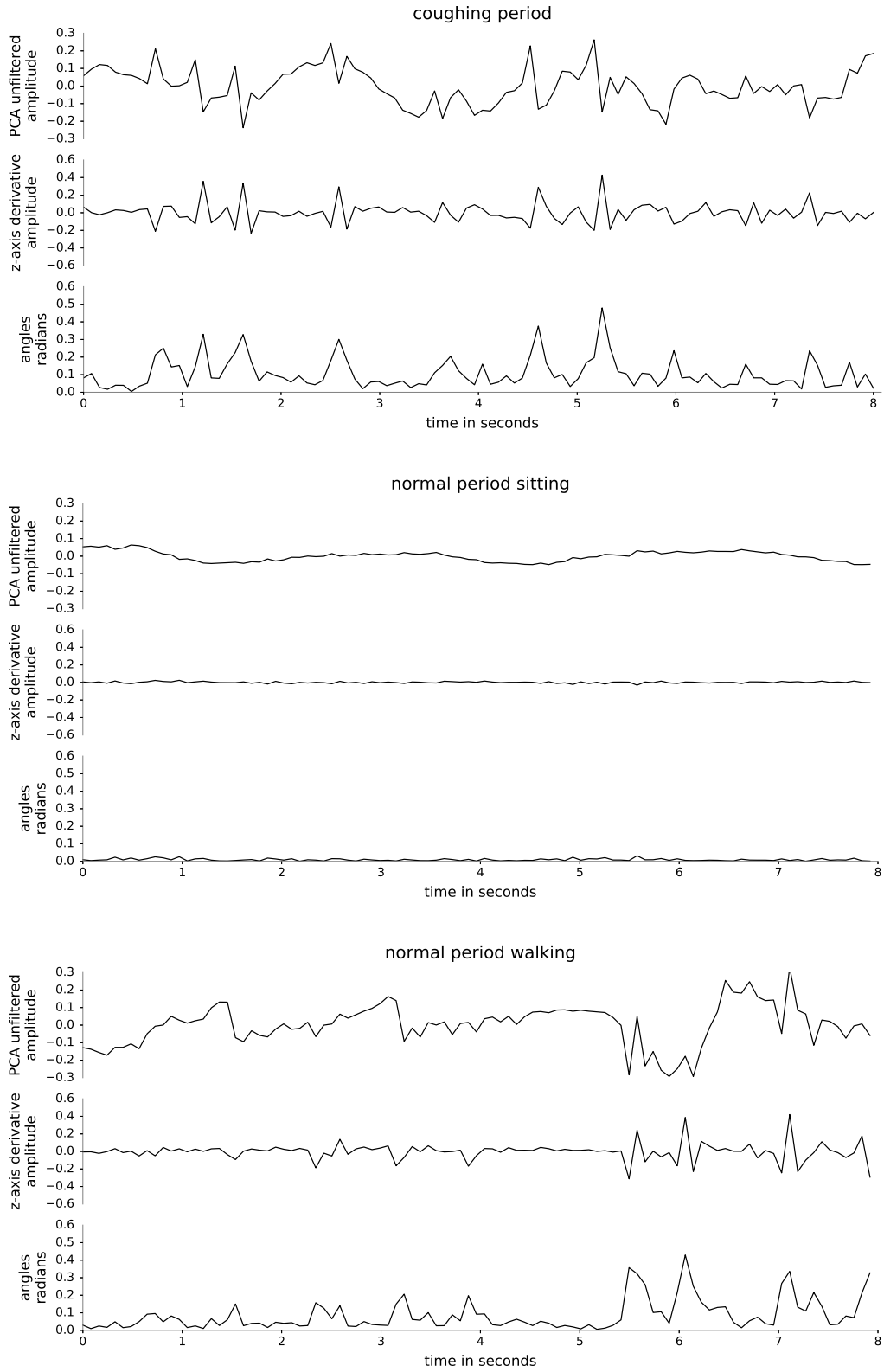

Figure 4.7: Different periods of 8 seconds, one containing coughs while sitting, two containing no coughs while sitting and walking.

already uses the angle between two consecutive vectors which has a similar effect as the derivative.

The derivative, as calculated by the difference between consecutive values in a discrete signal, can be repeatedly applied to the same signal, which is denoted as nth-order derivative. It might make sense to use a higher-order derivative in order to smooth non-spikes even more and thus different orders were tested for the classification task.

The remaining task is to capture the height and frequency of spikes into a simple measure calculated over the whole period. One approach is to take the variance of the spikes. As this results in a single value, the classification task is reduced to finding a threshold value for this variance. Another way is dividing the signal in a predetermined amplitude range into equal-sized bins and count the occurrences of the amplitude values in those bins, i.e. build a histogram of the amplitudes. The resulting histogram can then be fed into a classification model such as a neural network or an SVM. The SVM with a linear kernel was found to be superior to neural networks in all our tests, and therefore only the SVM results are presented.

In preparation for the classification task, the complete recording is split into windows of six seconds each. Six was chosen as the rounded mean length of all the coughing periods, leading to a total of 877 windows in the recording. Making the window size much smaller would mean that the amount of information available for the classification would get too small, whereas making it much longer would decrease the ability to localise the cough in time. Optimising the window size further would most likely lead to an overfitting on the data set, as the ideal window size depends on the duration of the coughs in the actual data and will be very specific to the subject and condition. A window was marked as containing a cough if it overlapped with at least two seconds of coughing as noted down by the subject.

For the variance threshold, all values in the range of the signal were tested with a precision of 0.0001. For the histogram, ranges between 0 and 0.3-0.7 led to the best results for all signals. The upper range limit was incrementally increased by 0.05. All ranges were tested with a number of bins between 5 and 55, in steps of 5. Again, these values were found to be sensible ranges for this experiment, as the performance of both classifiers decreased at the lower and upper limits for all signals. The SVM was trained and tested using 8-fold cross-validation.

Table 4.2 shows the results of our experiment, as measured by the F1-score:

$$\frac{2TP}{2TP+FP+FN}$$

where TP are the true positives, FP the false positives, and FN the false negatives. The F1-score is better suited than the accuracy metric for unequally distributed classes, which is the case here as there are a lot more non-coughing periods than coughing episodes. The best result was achieved with the SVM on the first order derivative of the z-signal with a F1-score of 0.92, or an accuracy of 0.989. The optimal histogram parameters for this result were 5 bins in a range of 0.0-0.5.

order	0	1st	2nd	3rd	0	1st	2nd	3rd
technique	var	var	var	var	SVM	SVM	SVM	SVM
PCA unfiltered diff	0.57	0.84	0.85	0.86	0.26	0.88	0.87	0.85
z-dimension diff	0.55	0.84	0.88	0.88	0.25	**0.92**	0.87	0.88
angles	0.82	0.81	0.78	0.75	0.85	0.82	0.78	0.77

Table 4.2: The results from classifying coughing periods represented by the F1-score.

In order to use this result for extracting features, the SVM is trained on the full data set with the best histogram parameters, and then saved as a model to be applied later on the study data.

When looking at the individual classifications in detail, it is evident that most of the FP stem from the walking periods. The technique therefore does not allow to separate walking from coughing, and should be used only when the subjects are sitting, i.e. during rest periods. As a preparation, the rest periods are split into windows of six seconds to equal the format of the training data and are then fed into the SVM. To mitigate the effect of false positives, a rest period is labelled as containing a cough if at least two 6-second windows in the period are labelled as coughing periods by the SVM.

4.2.6 Exercise execution

As mentioned previously, the movements during the exercises make it difficult to extract any breathing information from the sensor data. Instead, all the other movements should be analysed for helpful information.

This could, for instance, be the speed of execution of the exercises, as that would offer an estimation of how fit the patient is. As the exercises consist of periodical movements,

these should show up as frequency components in the frequency spectrum. The faster the subject conducts this motion, the higher the frequency will be. Additionally, the more regular the movements inside one period are, the more distinct and the higher the spike in the frequency spectrum. The spectrum should therefore allow deductions about the speed and regularity of the exercise movements.

Two spectra have already been discussed in Section 3.3: the Fourier spectrum (FS), and the marginal Hilbert spectrum (MHS). Both of them will be evaluated for this task. In order to reduce the dimensionality, the frequency spectra are split into equally sized bins and the mean of the amplitude values in each bin is stored. This way, the spectra can easily be fed into a neural network without having to use several hundred input units, which makes the training possible with less data and reduces the chance of overfitting.

As we are not interested in the breathing signal this time around and need a representation of the acceleration signals which is as close as possible to the raw data, the unfiltered PCA on the acceleration data has been chosen. Experiments with other signals discussed in Section 4.2.1 all show a clearly inferior result compared to the unfiltered PCA.

The parameters to be determined are the number of *Intrinsic Mode Functions* (IMFs, see Section 3.3) the signal is divided into, the cut-off frequency of the spectrum (i.e. what part of the spectrum is used) and the number of bins. As the exercise movements are very likely even slower than breathing, the lower cut-off value was always chosen to be zero. Instead of optimising these parameters on the data for the prediction task (and risk overfitting), they are instead optimised through another classification task, namely on classifying the exercises of the Wester Hailes study. This study was chosen as it very closely resembles the one in this project. As the Wester Hailes study did not entail logging the subject's well-being, the model cannot be directly optimised for this task. However, our hypothesis is that trying to distinguish between exercises will optimise for similar parameter values as it ensures that the frequency components specific to an exercise are emphasised.

In the Wester Hailes study, six patients performed a total of twenty-one exercise blocks with seven exercises each. Different upper cut-off values in the range between 0.5 and 6.0 Hz were tested, as well as bin counts ranging between 5 and 50. Figure 4.8 displays the F1 score for all parameter choices of upper cut-off value and bin count generated by two loops: the outer loop chooses the cut-off value, the inner loop the bin counts

which means that the small periodic patterns seen in the graph are due to the repeating bin counts. The exact parameter choices are less important than the fact that the MHS dominates the FS in all cases.

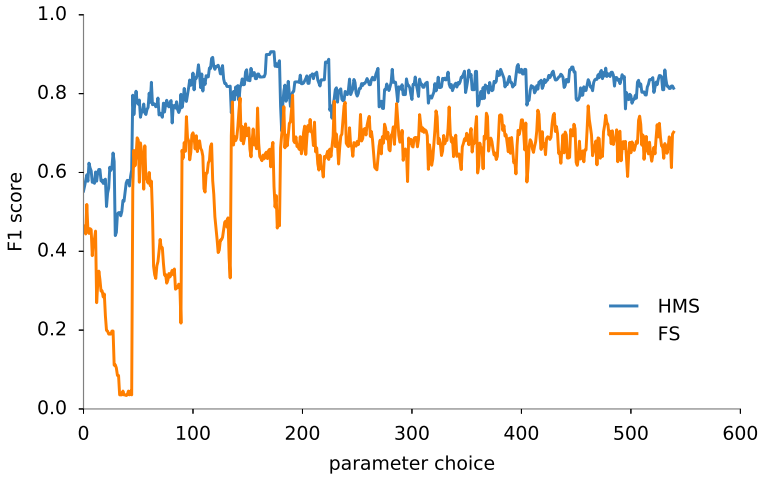

Figure 4.8: The F1 score for classifying exercise types with FS and MHS.

Figure 4.9 displays an alternative view on the results. The x-axis shows the maximum number of bins used for each run, and the y-axis shows the best F1 performance achieved in that run, when testing with all bin counts smaller than or equal to the maximum one. The reasoning underlying this is that the performance often does not increase if we use more bins and we are interested in a bin count which is as small as possible without compromising the performance.

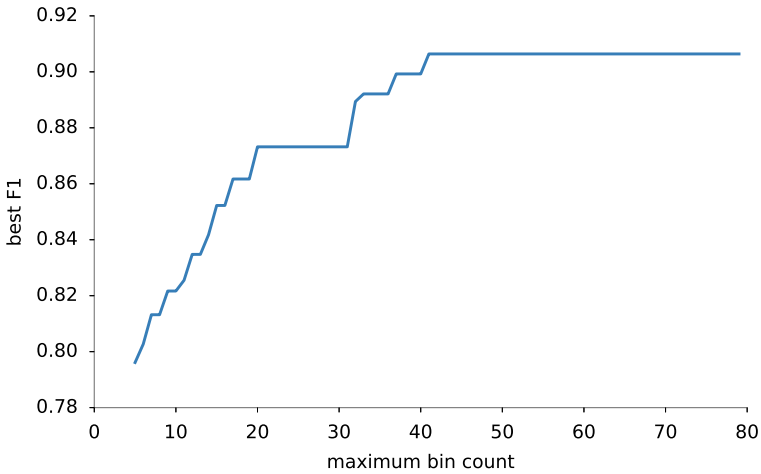

Figure 4.9: The best F1 score for different instances of maximum bin numbers with the MHS.

We observe that the improvements plateau at 41 bins where the best score of 0.906 was reached with a cut-off value of 2.0. Choosing 41 as the number of bins for our features would lead to a total feature count of $41 * 10 = 410$ features. Comparing these with the 24 features we have so far defined, one notes that the spectrum features are getting a disproportionately bigger weight. However, most of the 410 spectra features do seem to have a good predictive capability, as will be seen in Section 5.3.2.6, which means that the high proportion of those features is justified.

4.2.7 Other measures

Koley et al. [30] successfully applied additional features for detecting sleep apnea which were outside the scope of this study. These include other frequency spectra features which are all based on the power of frequencies in specific ranges or the proportion of the powers in different ranges, and also include complexity measures such as the *Approximate Entropy*.

The frequency spectra features have not been added as the ideal definition of the range endpoints are very much dependent on the task and the subjects. As there is no single known frequency range containing COPD-specific patterns, this would result in random guessing. Additionally, the MHS spectra during the exercises should already contain this information, especially as different bins of the spectrum will be combined inside the neural network during the experiments.

The complexity measures rely on the fact that a normal breathing signal is regular whereas irregularities indicate breathing problems. However, as already stated, there are no specific breathing irregularities we are looking for. Additionally, the breathing signals in Koley et al. were recorded in sleeping subjects which makes the form of the curve very regular. In recordings of moving subjects, the respiratory rate will vary over time and noise which is not related to breathing will be introduced through movements. The complexity measure would therefore most likely represent an alternative activity measure, although harder to interpret.

For these reasons it was decided to omit these other features from our analysis. This leads to a total of 434 features, 410 of them from the MHS frequency spectra and the remaining 24 from the time domain.

Chapter 5

Experiments and evaluation

Now that all the features have been discussed, it is time to extract these features from the current study in Sutton. Firstly, our main hypothesis regarding the results is stated, followed by a related hypothesis which is also analysed in the experiments. Afterwards, the feature extraction process is further elaborated and statistics about the study summarised. Lastly, the experiments will validate our main hypothesis, but also show the limitations of the current approach. They will also allow insights about pulmonary rehabilitation based on the importance of the features in the prediction task.

5.1 The hypothesis

The main hypothesis of this project is that it is possible to predict the well-being of a patient with COPD as measured by the CAT, based on the accelerometer data from the RESpeck worn during pulmonary rehabilitation exercises. In this context, "predicting" is meant in terms of estimating the CAT score as closely as possible with machine learning techniques on exercise data before the questionnaire entry.

Additionally, it is hypothesised that the progression of the disease for one single patient can be detected, i.e. that the accuracy of the prediction is good enough to detect trends.

The ultimate goal of the study would be to have a method of assessing patient well-being which requires no active involvement of the patient apart from wearing the RESpeck sensor during the pulmonary rehabilitation exercises.

5.2 Data preparation

During the writing of this dissertation, the study in Sutton was still active. The data analysed in the following is therefore only a current snapshot of the study over 133 days, from the 14th of March 2016, to the 25th of July 2016.

5.2.1 Diary entries

656 diary entries from 31 patients have been logged on the server during this time period, 84 of them without any selected answer (i.e. “I don’t know” for all questions), which have therefore been discarded.

One diary entry (from patient 4002) has been removed from the set because only one question was answered and the whole questionnaire was sent again immediately after that, which implies that the patient changed their mind about that one answer and then could only amend their choice by resending the whole CAT.

Figure 5.1 displays the CAT scores for all diary entries of six patients. Subjects 4002, 4019, and 4103 have realistic CAT scores in a narrow band which increases or decreases gradually, as would be expected, whereas subjects 4044, 4101 and 4136 submitted scores with an unrealistically high variance.

This finding suggests that some patients fill in the questionnaire with an absolute scale in mind, i.e. they assume values outside their typical score range are reserved for patients feeling considerably better or worse than themselves. Patient 4002 for instance would, at their current state, probably not fill in a value above 20, as there are patients in a clearly worse state whom they would expect to answer in that higher range.

Other patients, like the ones in the lower graph of Figure 5.1 seem to answer the questionnaire with a more relative scale in mind. 4044 for instance changes from the best (0) value to the second worst (39) and back to the best one in only a few days. This could either be an operating mistake of the questionnaire application, which seems unlikely given the otherwise mostly sensible answers, or it could mean that the subject simply compared his well-being to the previous days. If they felt considerably worse or better, they would score very high or low on each question, regardless of whether there are any other patients who might be feeling even worse/better.

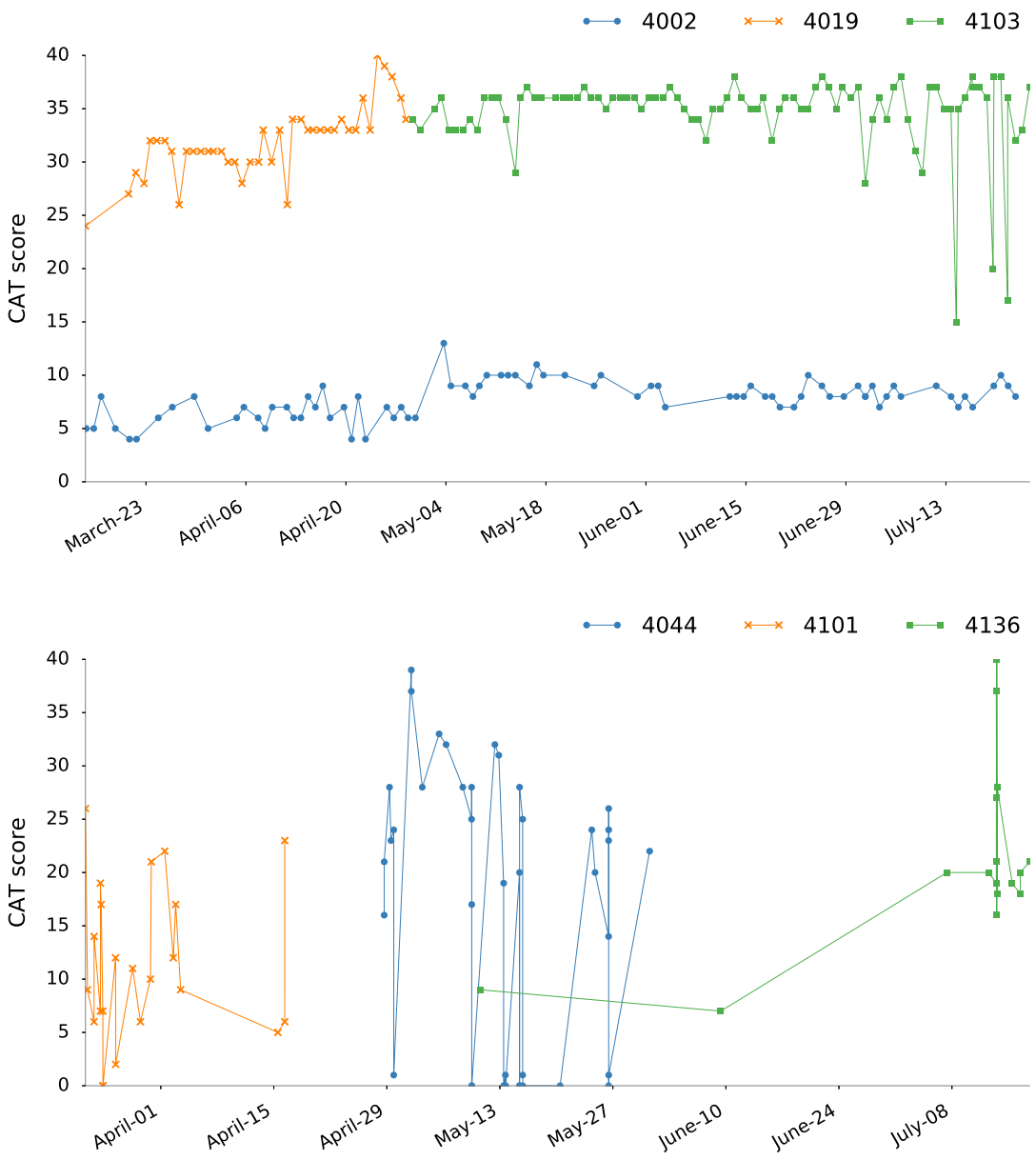

Figure 5.1: Distribution of CAT-scores for all diary entries of six patients.

This deviating pattern of answers can be seen as an inherent problem of the CAT, as the question statements are open to very subjective interpretation. “I cough all the time” could, for example, mean a cough every minute, every few minutes, or even only every hour if the patient has an otherwise healthy background.

This raises the problem as to how the different answer patterns are handled best for the prediction task. As we will see later, the patients with the more varying scores are also the ones who did the least amount of exercises. This basically annihilates the effect of the different answer patterns and we can therefore ignore the problem for the current

analysis. However, it should be kept in mind for future studies, where the CAT should be replaced by a questionnaire with more objective questions.

5.2.2 Determining valid exercise blocks

As the goal of the project is to estimate the CAT score from the rehabilitation exercises, each CAT entry has to be associated with exactly one exercise block. Naturally, this should be the most recent one before the entry and ideally, the questionnaires are filled in directly after the exercises. However, the study data shows that most patients do the exercises and diary entries independent of each another.

This poses the question how far apart in time the exercise block and the CAT entries should be at the maximum. A first intuitive value would be one day or 24 hours which, however, only leads to a total of 196 valid exercise blocks. Table 5.1 compares the amount of valid exercise blocks within the last 24 and 48 hours. The first row displays the count of exercise blocks which can be used for the prediction task.

If two diary entries are close to each other without an exercise block in between, they would get counted twice. Therefore, all occurrences after the first one are discarded (row number 2 in the table). If there are multiple exercise blocks before a diary entry, only the most recent one is taken.

	24h	**48h**
Valid exercise block, valid features	196	271
Valid exercise block, double count	9	25
Valid exercise block, but NaNs in features	5	6
No exercise period in the last x hours	268	166
Less than 10 exercise periods, none of them walking	18	18
At least 10 exercise periods, but less than 10 rest periods	1	0
More than 10 exercise and rest periods, but not the same number of them	0	8
At least one exercise was shorter than 5.5 seconds	13	24
No complete exercise block of 10 exercises, but a valid walking period	62	54
	572	572

Table 5.1: Statistics for the data extraction when looking at the previous 24 or 48 hours.

Some exercise blocks contain periods without any discernible breathing pattern, which lead to features with “Not a Number”-values (NaN). Replacing these NaN values with the mean value introduces errors in the training process later on which was only deemed

acceptable for the respiratory rates due to their importance. In the case of NaN values with other features, the whole exercise periods have been discarded from the prediction data, which in this case did not affect the total count by much.

The remaining rows in the table display other reasons why certain diary entries had no valid exercise block before them. Row number 4 shows that the number of diary entries which have no exercise period whatsoever in the the last 24 hours is a lot higher than for the 48 hours case. Switching to 48 hours therefore increases the number of valid exercise blocks by 75 or 38%. At the same time, the number of exercise blocks being counted twice also rises substantially. As there are only 318 complete exercise periods in total, the 271 extracted in the last 48 hours should provide a good trade-off between having enough data to train on and still being able to justify the time-relationship between exercise block and CAT entry.

In the following, the data from the past 48 hours is therefore used. Figure 5.2 again displays the statistics for 48 hours, but this time broken down by patient and with a higher-level view on the "invalid" diary entries. Note that the invalid blocks also include exercise periods where patients only walked, which did not help our analysis but counted as a clinically valid rehabilitation exercise in the study. This explains the high number of "invalid" exercise blocks for patient 4111 who only did the walking exercise most of the time. In total, 20 out of the 31 patients submitted diary entries with a preceding valid exercise block.

Figure 5.3 displays the CAT scores with valid exercises for all patients, i.e. the complete labels for our study. The affiliation of graphs to subjects and the progress of each individual graph is not important here. Even without this information, it is evident that the graphs from the lower plot of Figure 5.1 are largely missing. This is due to the fact that these three patients only contributed two valid exercise blocks in total.

Even more, after deleting every CAT score from the unprocessed diaries which seems improbable, i.e. all entries which deviate by more than 15 points from their neighbouring scores and even deleting complete diaries if the progress of scores seems very unlikely, the number of diary entries with valid exercises is only reduced by three. In other words, we do not have to worry about the improbable scores and which ones to exclude, but can simply leave them in the set.

Lastly, Figure 5.4 shows the same data as a histogram which better shows the general distribution of scores. These are the labels being trained on and being predicted in the

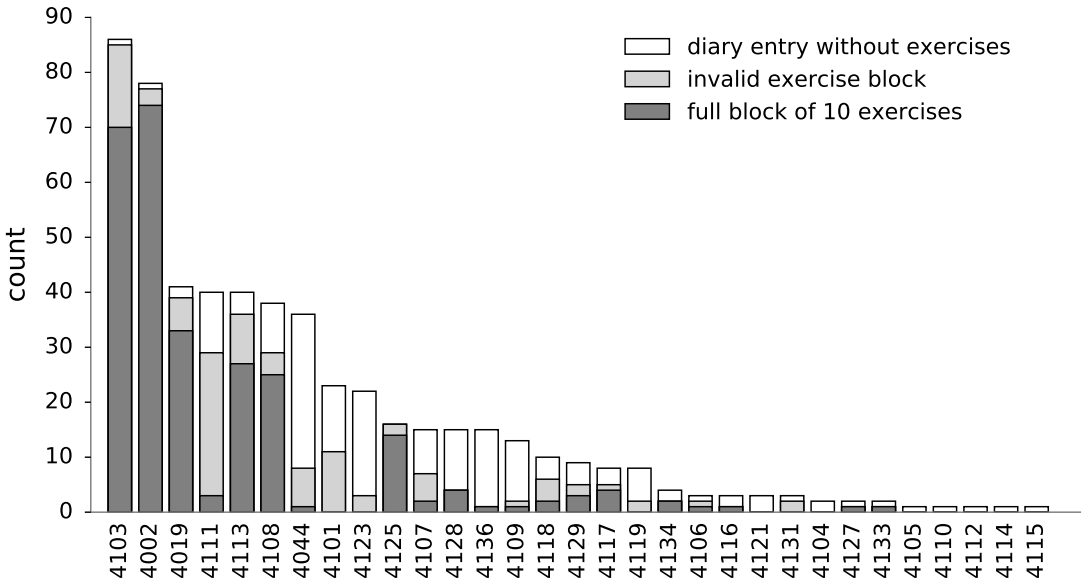

Figure 5.2: Statistics for each patient (denoted by patient id) when extracting from the past 48 hours before a diary entry.

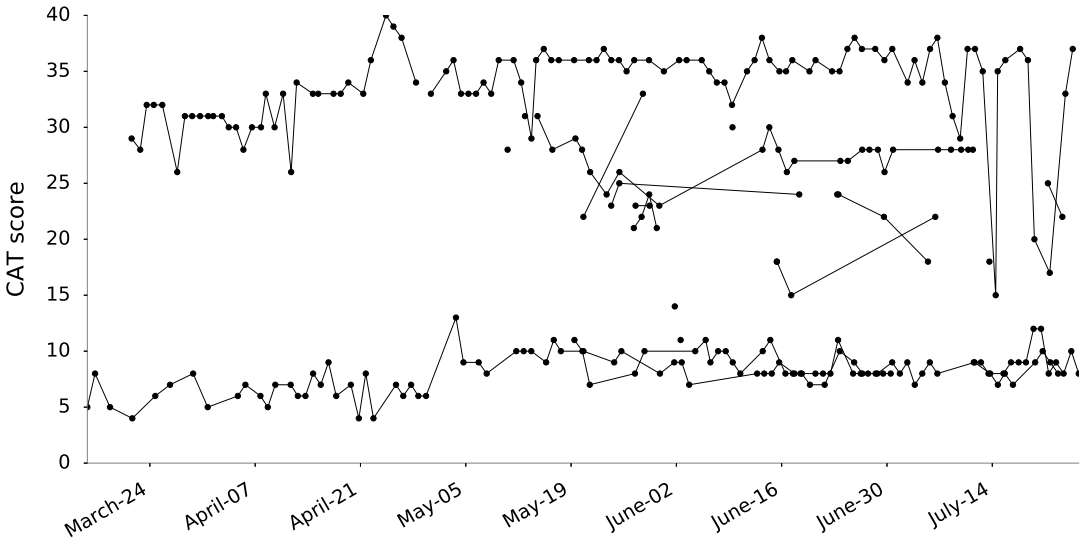

Figure 5.3: Distribution of all CAT scores which succeed valid exercises.

next sections. Note that for this study, there is a high prevalence of high and low values with a moderately represented middle ground. This is due to the agglomeration of most diary entries by only a few patients (cf. Figure 5.2) which are either at the healthier end of the scale, such as patients 4002 and 4113, or at the upper end, such as 4019 and 4103.

The lack of CAT scores below 5 can be explained by the fact that the study only includes patients who are already in treatment because of COPD and should therefore

have, by design of the questionnaire, a score above 0. A CAT score of around 0-15 is associated with stage 1-2 COPD ("mild" or "moderate") according to Tsiligianni et al. [44] which means that for patients belonging to this category, exacerbations are usually rare. Patients with scores from about 15 onward are in the more severe stages and are likely to have one or more exacerbations per year [21].

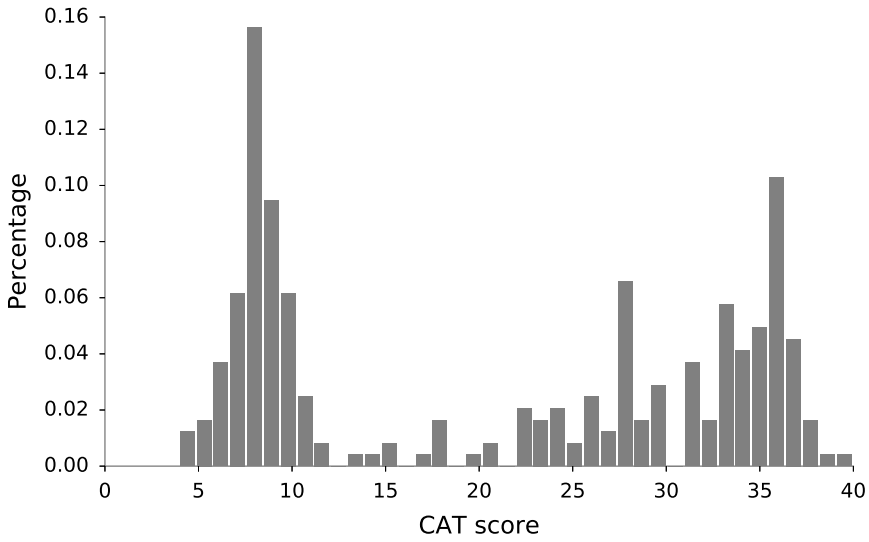

Figure 5.4: Distribution of CAT-scores for all diary entries succeeding a valid exercise block in 48 hours.

5.2.3 Removing erroneous measurements

When plotting the individual features against the CAT scores, an outlier can be detected for one instance in the *amplitude variance during rest* feature. This outlier is obviously above the expected deviation which means that there is likely a measuring error in the data for that exercise block. The whole instance should therefore be deleted, as it is quite likely that other features in that instance are affected by the error. Figure 5.5 shows the plot before and after removing the outlier.

In summary, the complete data for our prediction task is made up of 270 (271 minus one outlier instance) samples with 434 features each.

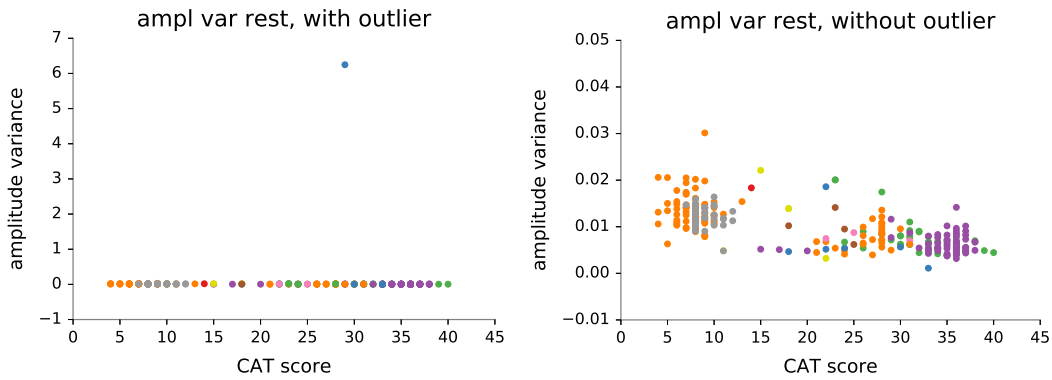

Figure 5.5: Plot of the *amplitude variance during rest* feature against the corresponding CAT scores, before and after removing the outlier. Different colours denote different patients.

5.3 Data analysis and experiments

Now that the data has been extracted, we can analyse the correlation between features and scores, and fit different regression models to the data. Before this, the baseline for the prediction task is introduced.

5.3.1 Baseline for prediction

There is to our best knowledge no other study comparable to this one, both concerning unsupervised respiratory monitoring of patients at home, and concerning the prediction of the patient's well-being based on the respiratory data. Considering this, there is no baseline performance in other pieces of research against which we can compare our predictions.

Instead, the minimum baseline is set to the error value which would result from always predicting the mean CAT score of all CAT scores in our data set. Two different error measures are used: the mean absolute error (MAE), as defined by the mean of the absolute deviation between the real and predicted values, and the root-mean-square error (RMSE), as defined by the root of the sum of each squared deviation:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |y^i_{pred} - y^i_{real}|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(y^{i}_{pred} - y^{i}_{real})^2}$$

The MAE weighs all deviations the same, whereas the RMSE weighs larger errors quadratically higher than smaller ones. For this reason, the MAE can be interpreted and evaluated directly and on its own, whereas the RMSE is better suited for comparing different approaches. Both both measures are reported for all the following results.

The baseline for the MAE is 11.34 and for the RMSE 12.03. With the MAE, this means that when predicting the mean CAT score (21.11) for all samples, we deviate on average by 11.34.

The mean, of course, is a very bad predictor and it should be easy to surpass this error if there is any relation between the acceleration data and the patient's health. A more challenging task is to surpass a linear regression model. Section 5.3.4 will establish this more advanced baseline.

However, we first investigated the expressiveness of each individual feature with the correlation coefficient.

5.3.2 Correlation coefficient of features and labels

An interesting question before training any model on our data is whether the features by themselves already contain any predictive power regarding the labels. This can be answered by calculating the Pearson correlation coefficient of each individual feature with the labels. Table 5.6 displays the coefficients for all features excluding the MHS spectra, as well as an extract from the top and lowest MHS spectra.

The coefficients range from -1 to +1, where -1 means perfect negative and +1 indicates perfect positive correlation. The features are therefore ranked by their absolute coefficient value, as both high negative and high positive values imply a strong predictive power. All coefficients from the non-spectra features apart from the worst one and all spectra features apart from the 95 lowest coefficients have a p-value below 0.05, meaning that the coefficient value is most likely not due to chance.

The intra-feature correlation is also high between the top-ranked, non-spectra features (0.58-0.96), although not too high to make them redundant.

feature	r	feature	r
angles rest	-0.84	MHS 6 - (0.29-0.34Hz)	-0.67
act measure rest	-0.81	MHS 9 - (0.83-0.88Hz)	0.64
lengths rest	0.78	MHS 3 - (0.00-0.05Hz)	0.63
mean bpms rest	0.72	MHS 9 - (0.78-0.83Hz)	0.63
mean volumes	-0.71	MHS 6 - (0.24-0.29Hz)	-0.63
ampl var rest	-0.69	MHS 9 - (0.88-0.93Hz)	0.62
variance bpms rest	0.66	MHS 9 - (1.85-1.90Hz)	-0.61
bpms rest 0	0.63	MHS 9 - (0.73-0.78Hz)	0.61
bpms rest 9	0.59	MHS 7 - (0.05-0.10Hz)	0.60
bpms rest 1	0.55	MHS 9 - (1.95-2.00Hz)	-0.59
bpms rest 8	0.54	MHS 4 - (1.51-1.56Hz)	0.59
bpms rest 6	0.53	MHS 9 - (0.93-0.98Hz)	0.59
bpms rest 4	0.49	MHS 9 - (1.90-1.95Hz)	-0.58
bpms rest 7	0.48	MHS 7 - (0.00-0.05Hz)	0.56
bpms rest 3	0.48	MHS 4 - (1.76-1.80Hz)	0.55
bpms rest 2	0.48	MHS 4 - (1.41-1.46Hz)	0.55
bpms rest 5	0.47	...	...
variance volumes	-0.39	MHS 5 - (0.24-0.29Hz)	0.02
angles exer	-0.38	MHS 6 - (1.27-1.32H	-0.01
act measure exer	-0.38	MHS 5 - (0.15-0.20Hz)	0.01
coughs count	-0.29	MHS 3 - (0.34-0.39Hz)	0.01
variance volume slopes	-0.26	MHS 0 - (0.10-0.15Hz)	0.01
ampl var exer	0.24	MHS 1 - (0.49-0.54H	-0.01
brr slopes	-0.23	MHS 0 - (0.15-0.20Hz)	0.00

Figure 5.6: Pearson correlation coefficients (r) of features with labels

5.3.2.1 Activity features

The highest coefficient value (-0.84) is achieved by the *mean angles during the rest periods*. As discussed in the previous chapter, large angles denote quick and large movements which can be viewed as a measure of the activity of the subject during the rest periods. A negative r-value means that the higher the angles, i.e. the bigger and faster the movements, the better the patient is feeling. Subjects are told to rest

and sit down in between exercises until they feel capable to start the next exercise. Apparently, healthier patients move around more during the rest periods. The coefficient value is surprisingly high – high enough, in fact, that it can predict the labels by itself with a MAE of 5.44 (RMSE 6.6) when fitting a linear regression model with 10-fold cross-validation on it, which is significantly better than the baseline.

Figure 5.7 shows the predicted values against the true values when fitting a linear regression model on the angle-rest-feature for all samples. We can see that the predictive power of just this single feature is surprisingly good on its own. Similar graphs can be displayed for all features with a high correlation coefficient.

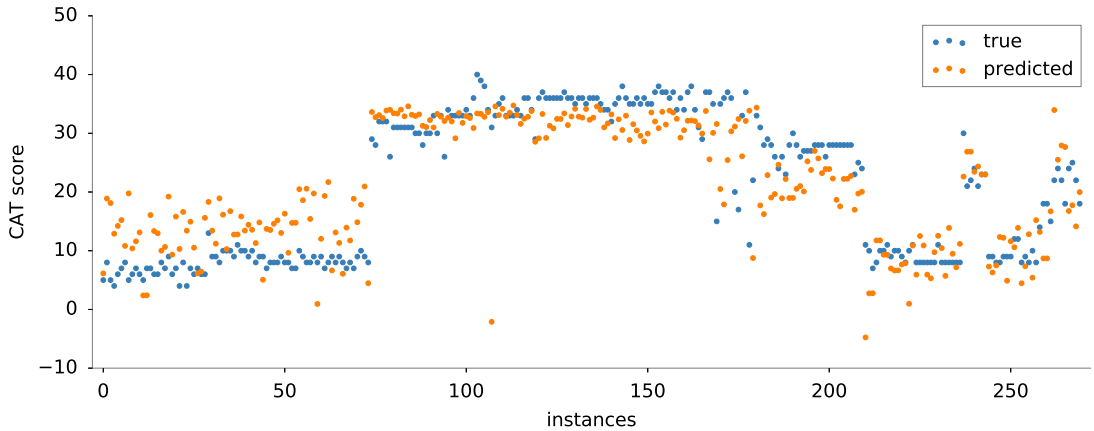

Figure 5.7: Actual and predicted scores when fitting a linear regression model to just the *mean angles during rest* feature.

This finding already confirms our main hypothesis when comparing with the worst-case baseline. Section 5.3.4 will establish the best baseline achievable with linear and polynomial regression models on all features. This *linear model baseline* will then be further improved by a non-linear neural network model. Before that, the implications of the other correlation coefficients are discussed.

The *activity measure* during the rest periods has the second highest r-value (-0.81). As can be seen from the very high correlation in between the angles and the activity measure (0.96 for rest, 0.98 for exercises), these two features seem to capture the same information. This agrees with our pooling of those features into the general "Activity features" category.

The third measure categorised as activity feature was the *variance of the amplitudes*. This feature also has a relatively high coefficient of -0.69, with a intra-feature correlation of 0.65 and 0.72 with the other two activity features (activity measure and angles

respectively). This suggests that the variance of the amplitudes has correctly been identified as an activity feature and that it is a suitable addition to the other two features.

The activity features during the exercise periods understandably have a considerably lower r-value than those during rest periods, as all subjects are moving around in this case. Still, a correlation of -0.38 for both the activity measure and the angles shows that healthier subjects typically perform larger or faster movements than subjects with a worse condition. This also corresponds to our intuition that healthier subjects conduct exercises more briskly.

5.3.2.2 Recovery features

The third highest coefficient – the *length of the rest periods* (0.78) – was categorised as a measure of breathing recovery in Section 4.2.2.2. The shorter the break between exercises, the healthier the patient seems to be. Interestingly, the *breathing recovery slopes* (brr slopes) which should more objectively capture the recovery time perform badly concerning the r-value (-0.23) and are even negatively correlated, meaning subjects with a more positive slope are less healthy, in contrast to intuition. This confirms the experiments on the Wester Hailes data set which also didn't show any clear connection between well-being and recovery slope. The reasons for this could be either that the method for computing the slopes is flawed, that the breath durations extracted from the RESpeck signal are not accurate enough, or that the measured slope is in fact correct, but not meaningful regarding the patient's health.

Nearly 60% of the recovery slopes are positive, as is the overall mean; a clear sign that the slopes are not only uncorrelated with health, but that they do not reflect what we would have expected, which seems to invalidate the slope calculation. However, in recordings after high intensity workouts and on a completely static subject during the rest period, the slopes show the clear negative trend we would expect (cf. Figure 4.4). The problem therefore lies either in the fact that a meaningful slope is only detectable after high intensity workouts and that the pulmonary rehabilitation exercises are not challenging enough, or that the subjects moved too much during the recording and therefore distorted the breath durations. A single outlier can change the slope from negative to positive which makes this feature particularly prone to recording errors.

The *length* feature on the other hand is basically immune to recording errors, as it is the only feature which is independent of the breathing signal, i.e. it can be measured

without the help of the accelerometer or any other respiratory monitoring device. On our dataset, a linear model trained and tested with 10-fold cross-validation leads to an MAE of 6.78, or RMSE of 7.69. However, as this feature can be consciously controlled by the subjects, it should not be used as an objective measure for the state of the health, at least not on its own.

5.3.2.3 Respiratory rate features

The mean respiratory rate measured in bpms has the next best coefficient (0.72). The higher the breathing rate during the rest periods, the worse the patient condition. This makes sense, as patients in a higher stage of COPD will have shallower breathing and therefore need to take in more breaths in the same time to compensate [21].

The mean over all rest periods seems to be a better predictor than the 10 individual mean breathing rates for each rest period, although the latter are already quite high, ranging between 0.47 and 0.63. The ranking of the bpms coefficients illustrates which exercises differentiate better between healthy and unhealthy subjects. These are not necessarily the most exhausting ones as measured by the mean respiratory rate over all periods, but rather those which are quite easy for healthier subjects but difficult for subjects in a worse state (cf. Table 5.2). Of all exercises listed in Section 3.1.1.2, “Sit-to-stand” seems to be the best differentiator in this sense, whereas “Shoulder press” is the worst.

The variance of the respiratory rates is also a good positive predictor (0.66), i.e. patients with a more varying respiratory rate score higher on the CAT. This might be explicable by changing levels of obstructiveness. Clearing his throat can suddenly enable a subject to take deeper and longer breaths, until the throat gets obstructed again.

5.3.2.4 Volume features

As already mentioned with the respiratory rates, COPD patients typically have shallower breathing than healthy persons as their airflow is limited by their obstructed airways. This also manifests itself in the high negative coefficient value of the mean volumes (-0.71): the higher the volume, the better (lower) the CAT score. This result is probably the one which can be most directly explained by the clinical definition of COPD.

Although this obvious explanation would validate our approach for extracting the volumes, and with that also validate both the accelerometer signal and the breathing

Exercise number	Activity	Mean RR	Rank coefficients
9	Step-ups	18.77	4
5	Bicep curl	18.7	6
7	Wall push-offs	18.3	5
8	Leg slide to the side	18.2	7
1	Sit-to-stand	18.06	1
4	Heel raises	17.98	9
6	Shoulder press	17.91	10
3	Squats	17.7	8
10	Walking	17.65	2
2	Knee extension	17.49	3

Table 5.2: Exercises ranked by mean breathing rate with the coefficient rank for comparison

signal calculation, it has to be considered with caution. High RESpeck amplitudes do not only arise due to voluminous breathing, but also due to other body movements. As we already know that healthier patients move around more during rest periods, the seemingly high tidal volumes could actually represent large movements.

With our current methods it is not possible to distinguish between irregular breathing and other low intensity body movements and it is questionable whether such a method exists, given the very diverse acceleration patterns of different subjects. A better solution to this problem is to ensure subjects sit during rest periods and are as static as possible. Although this cannot be guaranteed without supervision, the directions of the app could be made clearer in this regard.

The interpretation of the volume variance coefficient (-0.39) also depends on this factor. One the one hand, it could mean that patients with a higher variance of tidal volumes are in a better condition. This seems unlikely, however, given that a higher variance of breath durations is more strongly correlated with a *worse* condition. On the other hand, the variance could be skewed by movements. As we already analysed, healthier patients move around more which means that the variance will be higher due to the very large amplitudes. Although both explanations are possible, the second one is more likely due to the connection with the activity features.

As a reminder, the *volume slopes* are calculated by taking the slope of three consecutive volume values, which provides a more smoothed measure of volume changes. Even

though this measure worked quite well for Fekr et al. [18] for event detection, it does not seem to be meaningful for our task (-0.26). Just as the variance on the individual volume values, the coefficient is negative, which is not surprising given the close resemblance of the two measures.

5.3.2.5 Coughs count

The last of the non-spectra features is the cough count. The correlation value on this feature (-0.29) is a lot lower than expected given the good performance of the classifier on the separate recording in Section 4.2.5. In addition, the negative sign implies that more coughing correlates with a healthier state – a finding which seems obviously wrong, given that the CAT score is explicitly calculated based on the amount of coughing experienced.

Figure 5.8 shows the cough score extracted as feature, which denotes how many rest periods contain coughs, compared to the cough score from the CAT questionnaire, which is the answer of the first CAT question. The correlation of these two measures is -0.24 which is even slightly worse than when considering all questions.

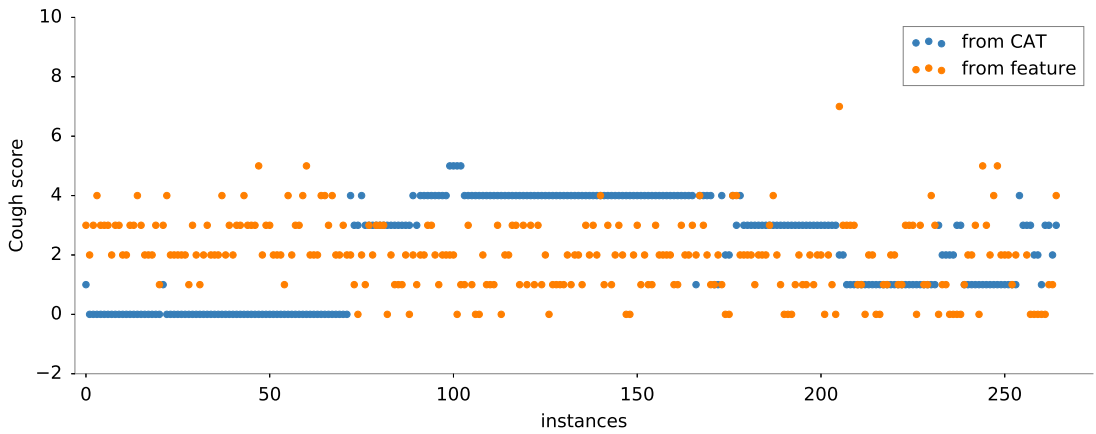

Figure 5.8: Cough scores from feature and from the CAT.

There are two possible explanations for this result: Firstly, the CAT measures the coughing in general and not only during exercise blocks. Therefore, most of the coughing will happen outside of the few short rest periods. The rest periods might then not be representative; even more, the subject is possibly coughing less during the exercise blocks as he might only start exercises during cough-free periods.

Secondly, subjects might move too much during the rest periods. In Section 4.2.5, we already noted that the coughing detection only works if the subject is static. Otherwise, the histograms of the amplitudes are more an alternative activity measure than a "cough detector". Interpreting the coughs as activity spikes would explain the negative correlation as we know that healthier subjects move around more during rest periods.

It should be investigated in future research whether the cough detection works in principle on other subjects, as long as they are mostly static, or if the detection method is dependent on the particular breathing pattern (amplitude) of one subject.

5.3.2.6 Spectra

In contrast to the relatively intuitive features discussed so far, the spectra are slightly more difficult to interpret. The coefficients of the spectra features range from -0.67 to 0.64. Each spectrum feature covers a frequency range of about $2.0/41 = 0.05$ Hz, which leads to 41 features between 0.0 and 2.0 Hz per exercise. A high positive r-value implies that patients with a bad condition tend to make movements in this frequency range, whereas frequency ranges with a high negative r-value are more likely covered by healthy subjects.

Figure 5.9 displays the r-value of all spectrum features as bar graphs partitioned by exercise. The y-axis denotes the r-value, while the x-axis shows the frequency range. Negative values have been coloured black to denote those frequency ranges which more often pertain to healthy subjects. This figure shows the coefficients in the range of 0.0 - 2.0 Hz as that was the range used for the features. As a reminder, this range was determined empirically in Section 4.2.6.

For the analysis below, it helps to also have the full range of frequencies up to 6.0 Hz (rounded from 6.25, half the sampling frequency of 12.5 Hz) which is therefore displayed in Figure 5.10. Note that Figure 5.9 is basically an extract of Figure 5.10.

In walking, the main frequency component is quite likely the walking speed due to the very periodic and regular nature of walking. The graph in Figure 5.10 suggests that the less healthy patients tend to walk at a speed of 0.5-1.0 Hz which translates to one step every 1-2 seconds. Healthy subjects walk faster with a growing coefficient up to 2.0 Hz, with a second peak at around 5-6 Hz which would correspond to a walking speed of 2 (5-6) steps every second. The highest negative coefficients are already captured in

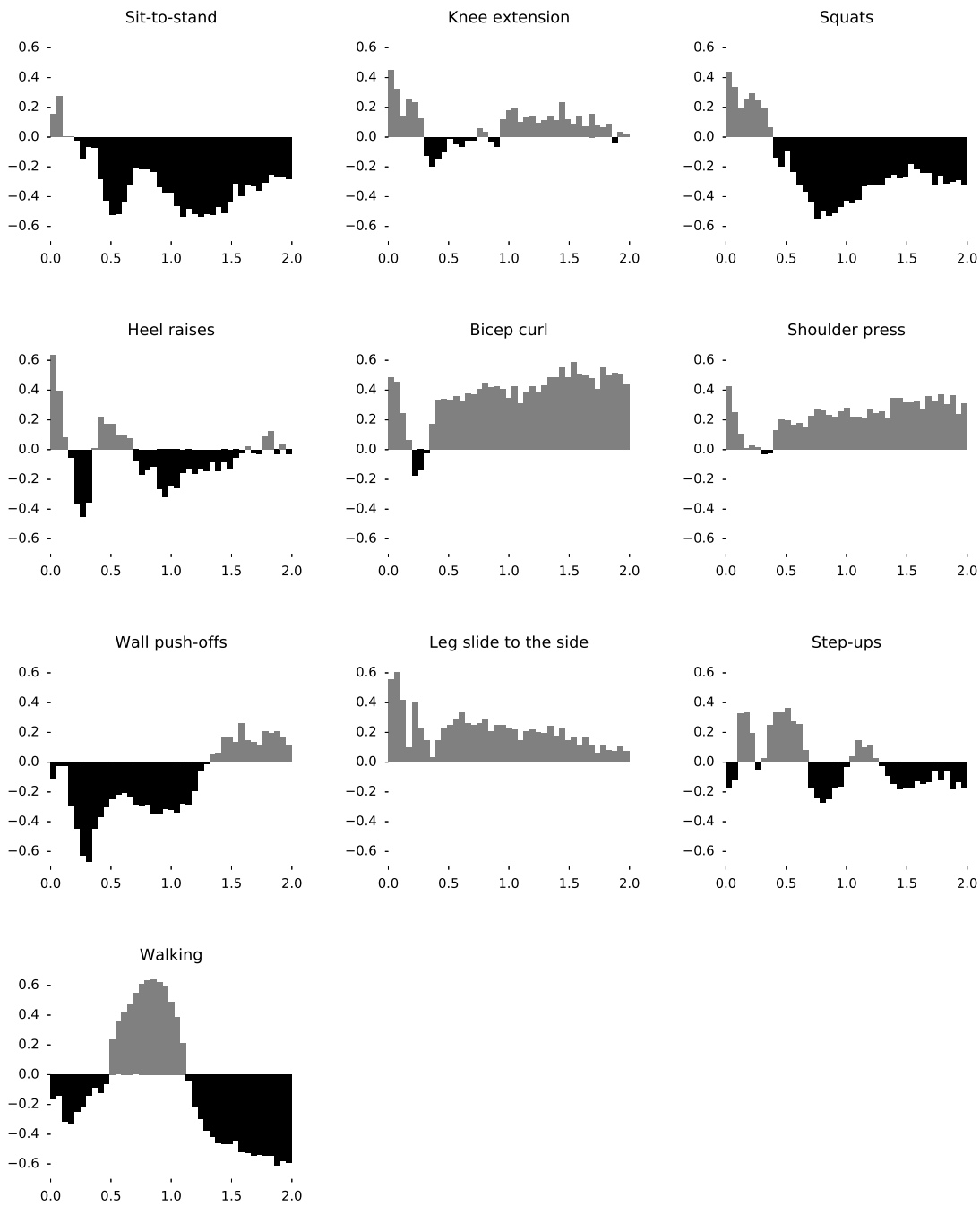

Figure 5.9: The correlation coefficients of the MHS displayed as bar graphs. The y-axis denotes the r-value, the x-axis the frequency (0.0-2.0 Hz). Negative values are coloured black, positive grey.

the smaller frequency range of 0.0 - 2.0 Hz, although the higher frequency components could still be meaningful for the analysis.

However, the coefficients alone are not enough to judge whether 2.0 Hz is indeed a suitable cut-off point. Figure 5.11 displays the mean power for each frequency

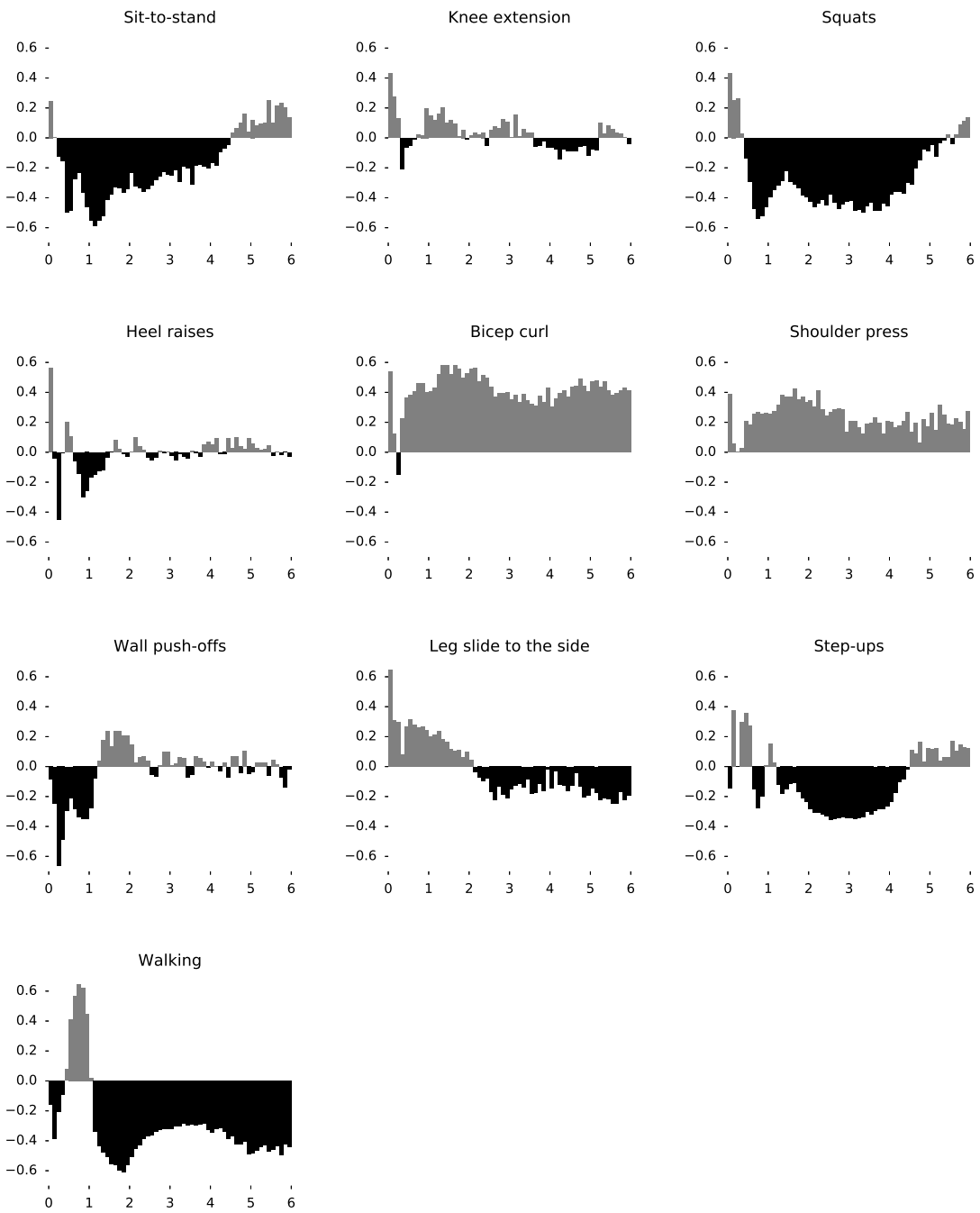

Figure 5.10: The correlation coefficients of the MHS displayed as bar graphs. The y-axis denotes the r-value, the x-axis the frequency (0.0-6.0 Hz). Negative values are coloured black, positive grey.

component calculated over all patients. A high power value means that most patients have this frequency component in their spectrum, whereas a value close to 0 implies that only few patients, if any, have this component. We can see that for most exercises, the power is around 0 for frequency values above 2.0 Hz. This means that large coefficients

above 2.0 Hz are likely based on very small changes in the frequency values and will therefore possibly generalise badly.

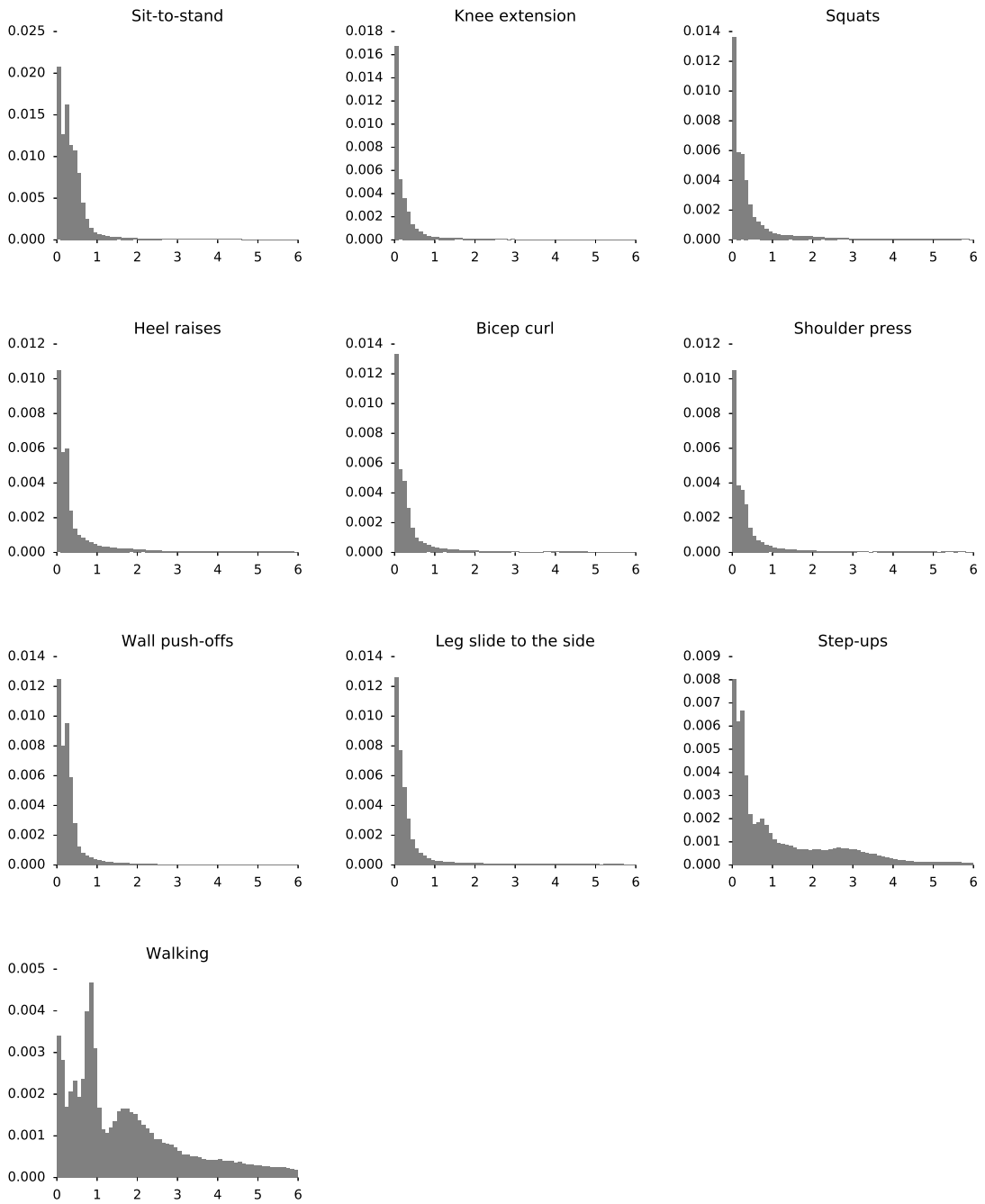

Figure 5.11: The mean frequency powers for all patients in a frequency range of 0.0 - 6.0 Hz (60 bins).

Sit-to-stand is an exercise whose coefficient graph is more difficult to interpret as there are very few positive coefficients. The graph suggests that the less healthy patients are in the lower range of 0.0 - 0.2 Hz. If we again interpret this frequency as depicting

the exercise movement, this would mean one stand-sit motion in around 5-10 seconds, which is realistic in the age group of the study.

Shoulder press, *Bicep curl* and *Leg slide to the side* all contain almost no negative coefficients in the range up to 2.0 Hz, and only the latter does have them in the higher frequencies. When focusing on the lower frequency range, this means that the prediction model will estimate those patients as healthier who have a low frequency value at points with a high positive coefficient. All three exercises contain no significant movement of the thorax/abdomen when performed correctly. Less healthy subjects possibly move more as they use other body parts to assist with the movements. With the Bicep curl, this could mean leaning back when contracting the biceps muscle to support the lift motion with the back/hip muscles.

Wall push-offs stand out from the graphs as the healthier patients here seem to do the exercises more slowly as opposed to faster than the less healthy patients. A possible explanation is that wall push-offs do not have a clear start and endpoint of the movement as most other exercises. This means less healthy subjects might be moving in a smaller range than healthy ones which allows them to complete the movements faster.

From the graphs, it can also be seen that some exercises are better able to differentiate between patient states in general. Table 5.3 displays the mean absolute coefficient per exercise, calculated on the 0.0 - 2.0 Hz range. Considering the full range up to 6.0 Hz changes the ranks only slightly. Walking is the best differentiator – a result which supports the use of this single exercise as its own valid exercise block, as has been done in the study. In future studies it might make sense to let patients only complete a subset of exercises which have a high mean coefficient. This would possibly decrease the prediction accuracy only slightly while increasing the completion rate as the whole exercise block will be shorter.

5.3.3 Linear regression

5.3.3.1 The models

Now that the meaningfulness of most features has been confirmed and a preliminary result of 5.64 (MAE) or 6.88 (RMSE) has been achieved using only the *mean angles during rest* feature, we will investigate how well we can predict the CAT scores using

Exercise	Mean abs. coefficient
Walking	0.397
Bicep curl	0.390
Sit-to-stand	0.322
Squats	0.308
Wall push-offs	0.237
Shoulder press	0.231
Leg slide to the side	0.215
Step-ups	0.166
Heel raises	0.155
Knee extension	0.119

Table 5.3: Mean absolute coefficients for each exercise

all features and different linear regression models. In preparation for the training, all features were scaled to have a mean of zero and standard deviation of one.

The most basic regression technique is the *Ordinary Least Squares (OLS) Regression*. The goal is to predict the labels, the CAT scores in our case, by a linear combination of each of the features. The deviation of the actual from the predicted value is squared and the sum of all squares taken as error value and the goal is to minimise this error.

A common problem when fitting machine learning models is *overfitting*, i.e. fitting the model to the training data too well, which means it will lose its capability to generalise. Linear regression models are prone to do this when the number of features exceeds the number of instances being trained on, as is the case here [23]. In linear models, this manifests as needlessly high coefficients which are then prone to slight changes in the test data.

In order to avoid this, regularisation factors are applied to keep the coefficients as low as possible. The L1 regularisation adds the absolute coefficient values to the error term, weighted by a *complexity parameter* α, whereas the L2 regularisation adds the squared coefficient values to the error term. The basic linear model extended by the L1 regularisation is called *Lasso Regression*, and with the L2 regularisation, the *Ridge Regression*. The *Elastic Net* includes both L1 and L2 [40].

In addition to constraining the the coefficient size, the L1 regularisation term in Lasso and Elastic Net often sets some of the coefficients to 0 (i.e., has a sparse solution) and therefore performs a subselection of the most important features.

Looking at the individual plots of features against scores (not displayed here due to space constraints as there are 434 plots in total), there are no evident polynomial relationships between the two. In addition, checking for polynomial connection involves generating even more features which will increase the overfitting problem. The following analysis therefore focuses on the above-mentioned, non-polynomial models.

5.3.3.2 Optimising the hyperparameters

In order to evaluate the best model and parameter combination, the data set is first split into a random training and test set, with 20% (54 samples) belonging to the test set.

The training set is then further split by 10-fold cross-validation, where the data is divided into 10 equally sized parts and each part is used once as a validation set. The validation is needed to find the best values for the hyperparameters of our models. OLS does not have any parameters, but ridge and lasso each have an α coefficient which denotes the importance of the regularisation term. Elastic net additionally has an *L1 ratio* parameter which determines the ratio of L1 and L2 regularisation.

RMSE is used to determine the best hyperparameter value for each model, as it penalises big deviations more which is sensible in this case. Figure 5.12 shows the RMSE values for different parameter values in Lasso, Ridge and Elastic Net. The minima are $\alpha = 0.36$ for Lasso, $\alpha = 106.5$ for Ridge and $\alpha = 0.55$, L1 ratio $= 0.0$ for Elastic Net. The parameters in Elastic Net show that the L1 regularisation does not improve the results independent of its weight.

Note that for a L1 ratio of 1.0, Elastic Net equals Lasso regression, and similarly for a L1 ratio of 0.0, it equals Ridge regression. Both Elastic Net and Ridge regression therefore calculated the same, best RMSE score of 4.61.

In Elastic Net and Lasso, the OLS term is divided by a factor dependent on the sample size which is not the case in Ridge regression. This explains the different scale of alpha values for these methods.

5.3.3.3 Comparing models and evaluating on test set

Table 5.4 displays the best MAE and RMSE value for each model on the training data with the best hyperparameter values from the last section. As elucidated, Ridge

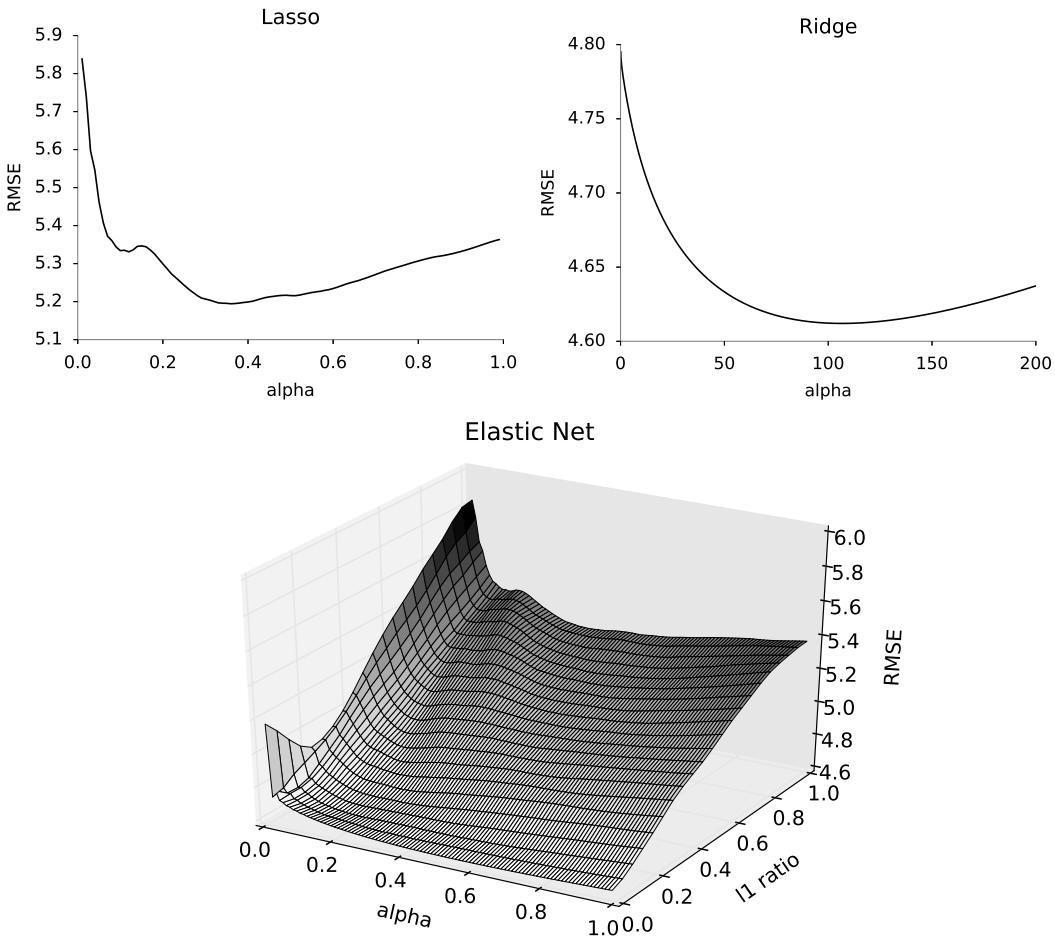

Figure 5.12: RMSE for different hyperparameter values and regression models.

Regression and Elastic Net calculate the same values with the optimal parameter settings.

For OLS, both scores have also been determined by 10-fold cross-validation. The extremely high values of those scores demonstrate the effect of overfitting when training without regularisation terms.

Model	MAE	RMSE	Parameters
OLS	2955527104.13	3958822894.18	-
Lasso	3.83	5.19	$\alpha = 0.36$
Ridge	3.53	4.61	$\alpha = 106.5$
Elastic Net	3.53	4.61	$\alpha = 0.55$, L1 ratio $= 0.0$

Table 5.4: Best MAE and RMSE for all four linear regression models

Ridge Regression and Elastic Net, therefore, have the best RMSE score. Evaluating the Ridge model with the optimal α value of 106.5 on the test set leads to a final MAE of 3.94, and an RMSE of 5.24. Both values are considerably better than the ones gained from OLS on the single best feature alone (see Section 5.3.2.1).

5.3.4 Regression with neural networks

Artificial Neural Networks (ANNs) are one of the commonly used machine learning techniques, used also in health care analysis (cf. [29, 45, 46]). Without any hidden layers, ANNs basically correspond to a linear model, but with one hidden layer, they can already represent any continuous function [24]. However, the exact configuration of layers and hyperparameters is very domain and problem specific.

5.3.4.1 Optimisation of the parameters

In our case, we only have $270 * 0.8 = 216$ training samples we can use for the 10-fold cross-validation (with a test set size of 20%), which is most likely too few to get a reliable optimum for the parameters. In order to smooth the error curves resulting from too few training samples, each cross-validation is repeated five times with different initial weight initialisations and the mean error of those runs stored.

For the hidden layers, the *Tanh* activation function was found to provide the best performance:

$$tanh(z) = \frac{\exp^{z} - \exp^{-z}}{\exp^{z} + \exp^{-z}}$$

Due to the regression task, the output layer only has one unit without any activation function, i.e. the output is a continuous real-valued number. The learning rate was set to 0.001 and fixed for all training epochs. No regularisation terms were applied as those only make sense with more complex networks and a bigger data set. The maximum number of training epochs was set to 500, but the training was stopped before, if the change in error was less than 0.001. The networks were programmed using the *scikit neural network* library in python [41].

Figure 5.13 displays the RMSE for a different number of units in the hidden layer, for the case that only one hidden layer is used. The error follows a clear trend, with the

lowest values at around 3-7 units. The minimum lies at 6 units with an RMSE of 4.766, which is worse than the cross-validation error on the Ridge Regression model.

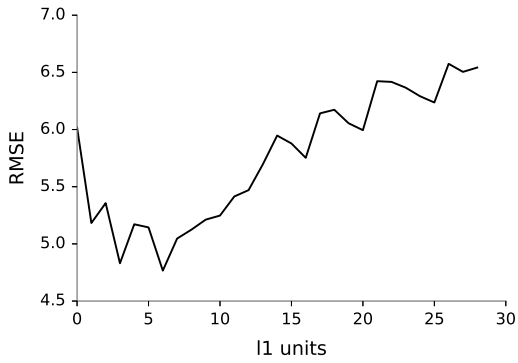

Figure 5.13: RMSE for different number of units in one hidden layer.

With two hidden layers (h1 and h2), the performance surpasses the linear models. Figure 5.14 shows the results of using different sizes for the first and second hidden layer in steps of five units. As evident from the graph, the errors follow a clear trend regarding the h1 size, although the error curve is not as smooth as with the Ridge Regression model.

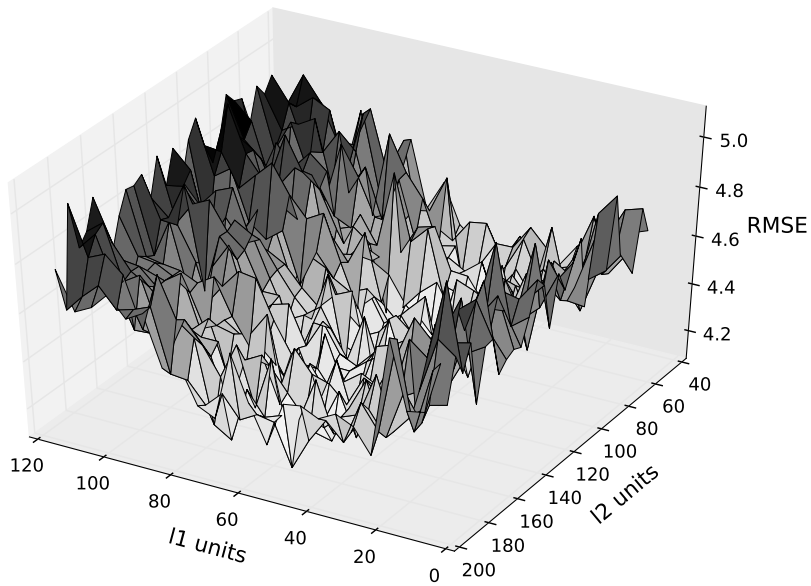

Figure 5.14: RMSE for different number of units in two hidden layers.

The global minimum lies at 40 units in h1, and 85 in h2, with an RMSE of 4.144. The h1 unit size is clearly the optimum for most h2 sizes, as can be confirmed by taking the

mean over all h2 values for each h1 value (Figure 5.15). The h2 values for h1_size = 40 are more erratic which means that 85 is not a clear optimum and might change with more training data.

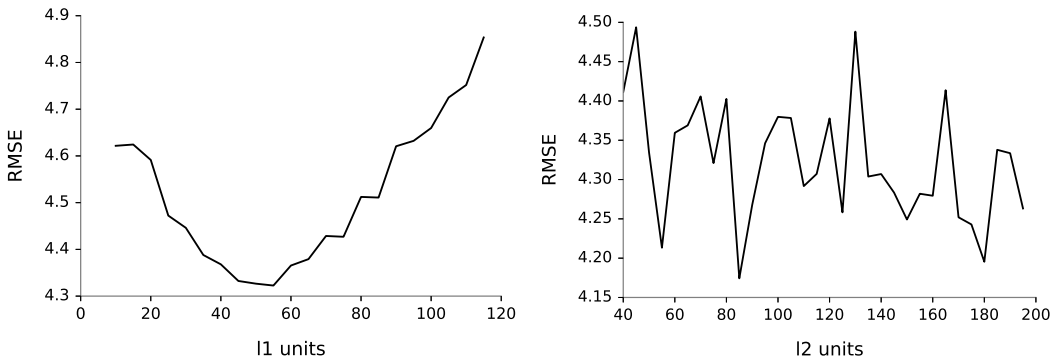

Figure 5.15: RMSE for the mean over all h2 values for each h1 value (left) and the h2 values for an h1 size of 40 (right).

With even more hidden layers, the performance is again worse than with Ridge Regression.

Applying the optimal two-layer network configuration to the test set and again running it five times, taking the mean of the error scores for a more reliable estimate, returns an RMSE of 4.98 and MAE of 3.119. This is already a lot better than the performance on the Ridge Regression model (RMSE 5.243, MAE 3.94). Note that the Ridge Regression model returns the same error value in each run.

One possibility to further optimise the performance is to run the network multiple times and to take the mean of the individual predictions as the final prediction (rather than the mean of the errors as before), a technique called *neural network ensemble* [22]. Figure 5.16 displays the RMSE of cross-validation on the training set for an increasing number of networks in the ensemble, all of which have the same "optimal" construction as determined above.

As can be seen from the graph, the performance increases considerably just with a few runs, reaching a minimum RMSE value of 3.7 with six networks. Testing the performance with six networks on the test set leads to an RMSE of 4.639 and MAE of 2.77 which is again a lot better than the performance with only one network. In the following analysis, we will therefore use a two-hidden-layer neural network with 40 units in h1, 85 in h2, and take the mean of 6 runs as the prediction.

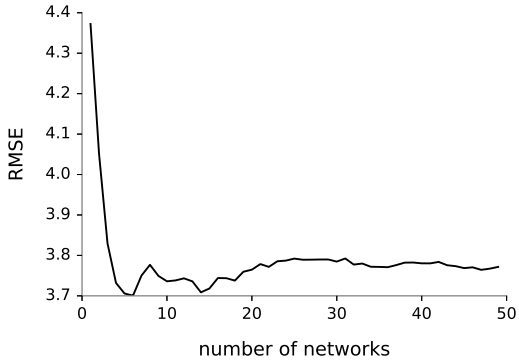

Figure 5.16: RMSE for different numbers of neural networks (*ensembles*) when using the mean prediction of the networks.

5.3.4.2 Prediction for each patient

So far, the data has been split into random subsets for training, validation and testing. This means that the data of one patient can be spread over all three sets, which allows the ANN to learn the characteristics of a patient and then assign those instances with similar values in the test set a score typical for that patient.

However, in clinical practice, the ability of the model to accurately identify the health of new, previously unseen patients is of greater interest. This ability can be examined by splitting the data by patients, training on the data of all but one patient and then testing on the remaining data.

Due to the skewed distribution of data amongst subjects (cf. Figure 5.2), this method will result in substantially varying sizes of the training sets. Leaving subject 4002 out of the training set leads to only 196 training samples, for instance, whereas leaving out patient 4044 will result in 269 training samples as the latter only has one valid exercise block. This difference means the results have to be compared and interpreted with caution, especially those of 4002 and 4103 which make up the biggest part of our data set.

Looking at the predicted vs. actual values for patient 4002 in Figure 5.17, the effect of this imbalanced distribution can be observed. As 4002 covers most of the lower-end scores, the model will only have few instances to learn the lower-end characteristics. It therefore assigns the features of 4002 in the test set a lower, even negative score. The negative scores can simply be set to zero, which still makes most of the predictions too

low, however. A better solution for future analysis would be to record a healthy subject in the age group of the patients and use his data as a lower limit of the model.

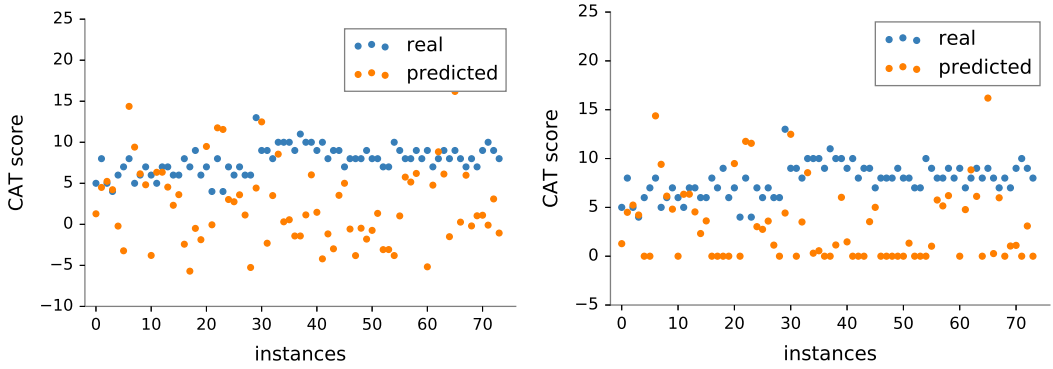

Figure 5.17: Predicted vs. actual scores for 4002. On the left with negative values, on the right without

Figure 5.18 shows the RMSE and MAE for each left-out subject with all negative values set to zero. All in all, the predictions match the actual scores really well, with a mean RMSE of 6.44 and a mean MAE of 6.096. Weighted by the number of instances per patient, the errors go up to 7.273 for the RMSE and 7.04 for the MAE. The scores are close to each other as they are the same for patients with only one valid instance. For some patients, the predictions seems to be a lot more accurate than for others. Figure 5.19 displays the RMSE for each patient (MAE is very similar).

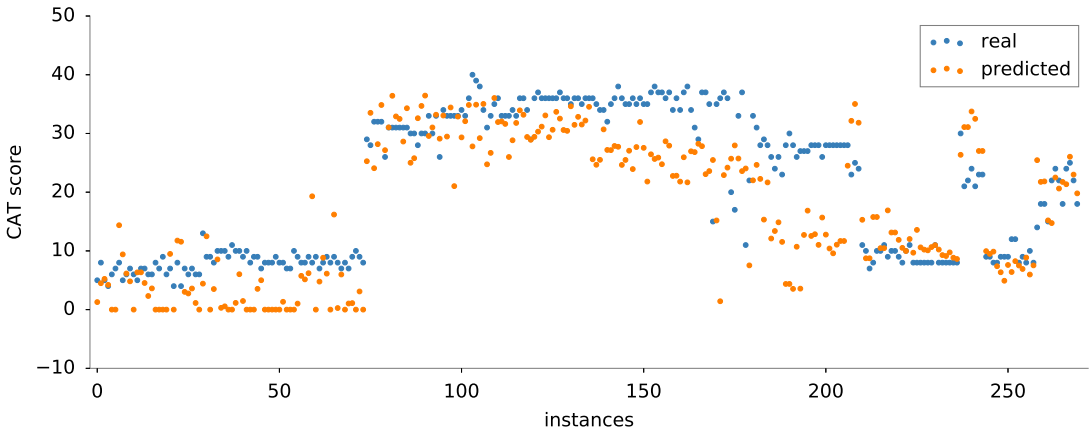

Figure 5.18: Predicted vs. actual scores with an ANN for all patients.

As we can see, the predictions for a few patients are far more off than for others. Interestingly, these are mainly the patients with only 1-2 instances, such as 4106, 4107 and 4127. 4108 falls out of this pattern with 25 instances. Apparently, this subject had a scoring scheme which differed from the other subjects substantially.

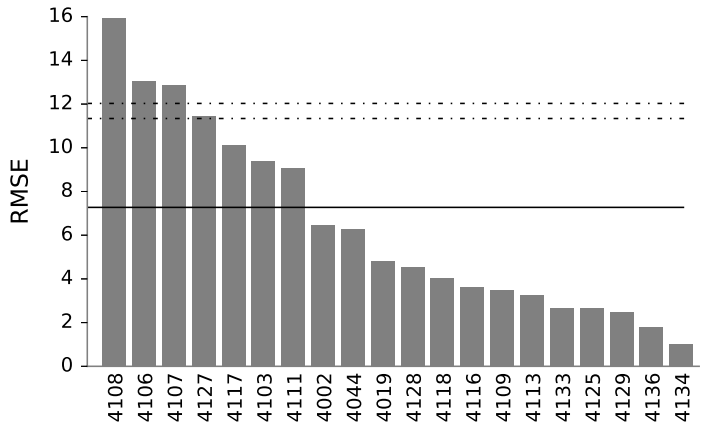

Figure 5.19: RMSE for each patient. The dotted lines mark the baselines, the continuous one the mean RMSE.

Looking back at Figure 5.18, the general performance of the regression models should be increased by collecting more data for the mid-range and edge-cases of scores, as this will more clearly define the boundaries between scores. Having more data with labels of around 20, for instance, will possibly push the predictions at around 150 instances up into a smaller and more accurate range.

Figure 5.20 displays the same graphic as before, but for Ridge Regression. Interestingly, the mean accuracy over all patients is better than that of the ANN when not weighted by the number of instances (RMSE 6.440, MAE 6.096). However, as can be seen from the figure, the predictions for some patients are off by a wide margin, in particular 4002, which leads to an RMSE of 11.646 and MAE of 11.283 when the weighted average is taken.

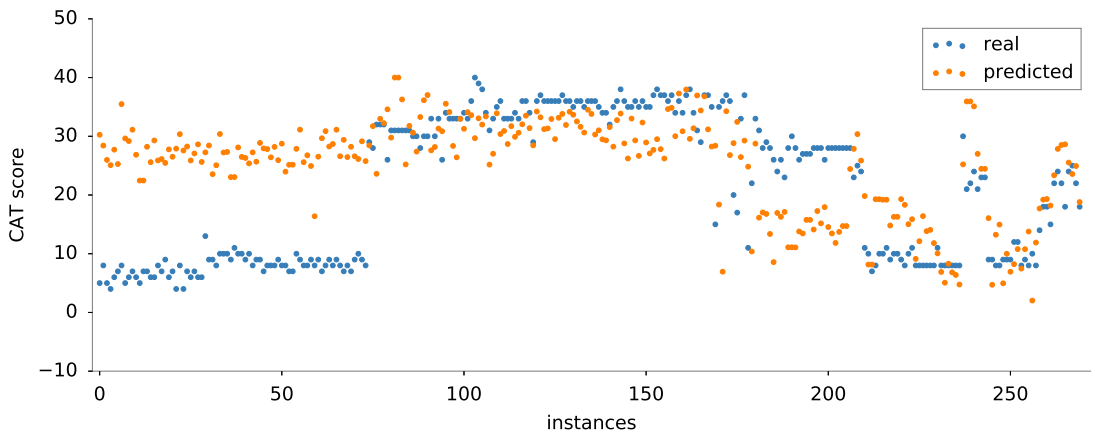

Figure 5.20: Predicted vs. actual scores with Ridge Regression for all patients.

5.3.5 Analysing trends for one patient

The final open question of our thesis is the soundness of our second hypothesis, namely whether trends in individual patients can be noticed and exacerbations detected.

One subject who shows a clear trend of scores is 4002 (see Figure 5.17. After 28 instances, the patient logged a score of 13 which is considerably worse than the score 6 in the previous entry. Even though we do not have the assessment of a physician for this incident, it could be called an exacerbation based on the definition of "an acute worsening of the condition" [21]. After this exacerbation, the mean of the scores increased by 2.4.

Ideally, we would be able to predict this worsening of the condition with the ANN. However, as evident from the graph, the trend is not detectable in the predicted values when the training is done on the other patients. Even more, the mean of the first 28 scores in the predicted values is higher rather than smaller than in the subsequent instances.

If we train both the ANN and Ridge Regression model on the first 28 instances of 4002 and all other patients, and test on the remaining instances of 4002, we get the plots in Figure 5.21.

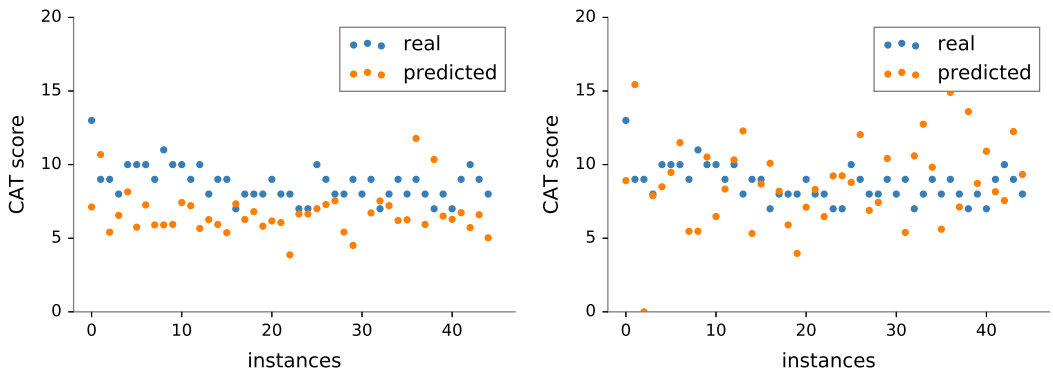

Figure 5.21: Predicted vs. actual scores with the ANN (left) and Ridge Regression model (right) for the second half of instances of 4002.

The ANN has the lower RMSE and MAE values (3.09 and 2.616, compared to 3.497 and 2.688), but the Ridge Regression model better captures the rising trend as its mean (8.951) is close to the mean of the actual values (8.71), whereas that of the ANN is closer to the first half of instances of 4002 (6.888). However, neither prediction is close enough to be able to support our hypothesis regarding the trend in individual patients.

A possible explanation why the ANN predictions did not rise in the test set is that the scoring pattern of 4002 might have changed, even though the actual well-being did not. This could be explained by the fact that unpleasant feelings are remembered better than pleasant ones [7] and that the worsening of the condition is thus experienced more strongly than the subsequent improvement. If the patient takes his past scores into account when answering the CAT, the score will therefore improve less than it deteriorated.

Unfortunately, trends in other patient's data similarly cannot be predicted accurately enough by either model, as can be observed from Figure 5.20 and 5.18. For this project, we have to conclude that either our models are not good enough to capture fluctuations in the individual patient's health, or that the accelerometer data just does not contain this information. The very subjective nature of the CAT, the time-difference between exercise block and diary entry and, lastly, the imprecision of the recording method with the accelerometer all add to the impreciseness of the predictions.

Chapter 6

Conclusion and future work

This thesis presented a novel way to assess the well-being of patients with COPD. The analysis was performed on a study with 31 COPD patients who completed pulmonary rehabilitation exercises at home while wearing an accelerometer sensor (RESpeck) on their chest. In addition to the exercises, the patients regularly filled in the COPD Assessment Test, a questionnaire designed to estimate the patient's stage of COPD, according to the GOLD guide [21].

After having introduced and validated a new technique to generate the breathing signal from the acceleration data, numerous features in time and frequency domain were extracted such as the breathing rate, the tidal volume, the coughing frequency, and measures of activity. Using these features and an Artificial Neural Network with two hidden layers, the CAT score could be predicted with an RMSE of 4.639 (MAE of 2.77) which is much better than the baseline of 12.03 (11.34).

When testing one previously unseen patients, the RMSE rises to 7.27 (MAE 7.04). This is mainly due to the fact that our study data is dominated by a few patients and the training set size therefore plummets when leaving out those patients.

A new method for detecting coughs in the RESpeck signal was successfully validated on a separate recording, but proved to be unreliable for the actual patient data. Similarly, the recovery of the breathing rate in rest periods as measured by a linear regression slope did not show predictive capabilities in our study.

All other features had a very good correlation with the CAT scores, however. In particular, the height of the respiratory rate and the tidal volumes are both not only

highly correlated, but can also be measured by other respiratory monitoring devices which makes our results applicable for clinical practice.

To the best of our knowledge, this was the first work analysing long periods of respiratory data in order to assess the general well-being of a patient, instead of only detecting specific breathing disorders. The predictions gained with the ANN are surprisingly good given both the subjective nature of the questionnaire and the noisy recordings from the RESpeck.

In order to further optimise the predictions, data from more diverse patients, especially for the edge-cases and the middle ground of the scores should be collected. This might allow the predictions to show the trends in individual patient's well-being and possibly foresee exacerbations.

The tidal volumes extracted from the RESpeck should be validated with a spirometer so that they can be used to make clinically proven statements about the connection between long-term volume recordings and breathing disorders.

Lastly, the ultimate goal of this research would be a functioning real-time alert system for patients and physicians based on the unsupervised recording of the RESpeck sensor.

Bibliography

[1] Daniel Alvarez, Roberto Hornero, J Víctor Marcos, and Félix del Campo. "Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis." In: *IEEE transactions on bio-medical engineering* 57.12 (2010), pp. 2816–2824.

[2] Donald H Arnold, David M Spiro, Renee' a Desmond, and James S Hagood. "Estimation of airway obstruction using oximeter plethysmograph waveform data." In: *Respiratory research* 6 (2005), p. 65.

[3] Mahdi Jan Baemani, Amirhasan Monadjemi, and Payman Moallem. "Detection of respiratory abnormalities using artificial neural networks". In: *Journal of Computer Science* 4.8 (2008), pp. 663–667.

[4] Sean J. Barbour, Christine A. Vandebeek, and J. Mark Ansermino. "Increased tidal volume variability in children is a better marker of opioid-induced respiratory depression than decreased respiratory rate". In: *Journal of Clinical Monitoring and Computing* 18.3 (2004), pp. 171–178.

[5] A. Bates, M. J. Ling, J. Mann, and D. K. Arvind. "Respiratory Rate and Flow Waveform Estimation from Tri-axial Accelerometer Data". In: *2010 International Conference on Body Sensor Networks* 191.1 (2010), pp. 144–150.

[6] Andrew Bates, D. K. Arvind, and Janek Mann. "Wireless monitoring of post-operative respiratory complications". In: *Proceedings of the 2nd Conference on Wireless Health - WH '11* (2011), pp. 1–9.

[7] Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. "Bad is stronger than good." In: *Review of General Psychology* 5.4 (2001), pp. 323–370.

[8] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007, pp. 561–569.

[9] Nan Bu, Naohiro Ueno, and Osamu Fukuda. "Monitoring of respiration and heartbeat during sleep using a flexible piezoelectric film sensor and empirical mode decomposition". In: *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings* (2007), pp. 1362–1366.

[10] S. Burge and J.A. Wedzicha. "COPD exacerbations: definitions and classifications." In: *The European respiratory journal* 41 (2003), pp. 46–53.

[11] R. M. Cherniack and M. B. Raber. "Normal Standards for Ventilatory Function Using an Automated Wedge Spirometer". In: *American Review of Respiratory Disease* 106.1 (1972), pp. 38–46.

[12] C R Cole, E H Ugene B Lackstone, F J Redric P Ashkow, C E Laire S Nader, M S Ichael L Auer, and A Bstract. "Heart-Rate Recovery Immediately After Exercise As A Predictor Of Mortality". In: *The New England Journal of Medicine* 341.18 (1999), pp. 1351–1357.

[13] Phil Corbishley and Esther Rodríguez-Villegas. "Breathing detection: Towards a miniaturized, wearable, battery-operated monitoring system". In: *IEEE Transactions on Biomedical Engineering* 55.1 (2008), pp. 196–204.

[14] C Davis, A Mazzolini, and D Murphy. "A new fibre optic sensor for respiratory monitoring". In: *Australasian physical & engineering sciences in medicine / supported by the Australasian College of Physical Scientists in Medicine and the Australasian Association of Physical Sciences in Medicine* 20.4 (1997), pp. 214–219.

[15] Gavin C. Donaldson, Tom M A Wilkinson, John R. Hurst, Wayomi R. Perera, and Jadwiga A. Wedzicha. "Exacerbations and time spent outdoors in chronic obstructive pulmonary disease". In: *American Journal of Respiratory and Critical Care Medicine* 171.5 (2005), pp. 446–452.

[16] G. B. Drummond, A. Bates, J. Mann, and DK Arvind. "Validation of a new non-invasive automatic monitor of respiratory rate for postoperative subjects." In: *British Journal of Anaesthesia* 107.3 (2011), pp. 462–429.

[17] Atena Roshan Fekr, Majid Janidarmian, Katarzyna Radecka, and Zeljko Zilic. "Respiration Disorders Classification With Informative Features for m-Health Applications". In: *IEEE Journal of Biomedical and Health Informatics* 20.3 (2016), pp. 733–747.

[18] Atena Roshan Fekr, Katarzyna Radecka, and Zeljko Zilic. "Design and Evaluation of an Intelligent Remote Tidal Volume Variability Monitoring System in

E-Health Applications". In: *IEEE Journal of Biomedical and Health Informatics* 19.5 (2015), pp. 1532–1548.

[19] M Folke, L Cernerud, M Ekström, and B Hök. "Critical review of non-invasive respiratory monitoring in medical care". In: *Medical & Biological Engineering & Computing* 41.4 (2003), pp. 377–83.

[20] GlaxoSmithKline. *COPD Assessment Test*. URL: `http://www.catestonline.org/english/indexEN.htm` (visited on 07/20/2016).

[21] *Global Strategy for the Diagnosis, Management and Prevention of Chronic Obstructive Pulmonary Disease*. Tech. rep. Global Initiative for Chronic Obstructive Lung Disease, 2016, pp. 1–111.

[22] Lars Kai Hansen and Peter Salamon. "Neural network ensembles". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.10 (1990), pp. 993–1001.

[23] Douglas M. Hawkins. "The Problem of Overfitting". In: *Journal of Chemical Information and Computer Sciences* 44.1 (2004), pp. 1–12.

[24] Jeff Heaton. *Introduction to Neural Networks with Java*. Heaton Research, 2008.

[25] Jean-Jacques Hosselet, Robert G. Norman, Indu Ayappa, and David M. Rapoport. "Detection of flow limitation with a nasal cannula/pressure transducer system." In: *American journal of respiratory and critical care medicine* 157.5 (1998), pp. 1461–1467.

[26] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. "The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis". In: *Proceedings: Mathematical, Physical and Engineering Sciences* 454.1971 (1996), pp. 903–995.

[27] P. W. Jones, G. Harding, P. Berry, I. Wiklund, W. H. Chen, and N. Kline Leidy. "Development and first validation of the COPD Assessment Test". In: *European Respiratory Journal* (2009).

[28] Romain Kessler, Elisabeth Stáhl, Claus Vogelmeier, John Haughney, Elyse Trudeau, Claes Göran Löfdahl, and Martyn R. Partridge. "Patient understanding, detection experience of COPD exacerbations: An observational, interview-based study". In: *Chest* 130.1 (2006), pp. 133–142.

[29] Bethany R. Knorr, Susan P. McGrath, and George T. Blike. "Using a generalized neural network to identify airway obstructions in anesthetized patients post-operatively based on photoplethysmography". In: *Annual International Con-*

ference of the IEEE Engineering in Medicine and Biology - Proceedings (2006), pp. 6765–6768.

[30] Bijoy Laxmi Koley, Member IEEE, and Debangshu Dey. "Real-Time Adaptive Apnea and Hypopnea Event Detection Methodology for Point-of-Care Applications". In: *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING* 60.12 (2013), pp. 3354–3363.

[31] L. B. Koley and Debangshu Dey. "Selection of features for detection of obstructive sleep apnea events". In: *2012 Annual IEEE India Conference, INDICON 2012* (2012), pp. 991–996.

[32] Barry J Make, Göran Eriksson, Peter M Calverley, Christine R Jenkins, Dirkje S Postma, Stefan Peterson, Ollie Östlund, and Antonio Anzueto. "A score to predict short-term risk of COPD exacerbations (SCOPEX)." In: *International journal of chronic obstructive pulmonary disease* 10.1 (Jan. 2015), pp. 201–209.

[33] J. Mann, R. Rabinovich, A. Bates, S. Giavedoni, W. MacNee, and D.K. Arvind. "Simultaneous Activity and Respiratory Monitoring Using an Accelerometer". In: *2011 International Conference on Body Sensor Networks* (2011), pp. 139–143.

[34] Jose M. Marin, Santiago J. Carrizo, Ciro Casanova, Pablo Martinez-Camblor, Joan B. Soriano, Alvar G N Agusti, and Bartolome R. Celli. "Prediction of risk of COPD exacerbations by the BODE index". In: *Respiratory Medicine* 103 (2009), pp. 373–378.

[35] Colin D. Mathers and Dejan Loncar. "Projections of global mortality and burden of disease from 2002 to 2030". In: *PLoS Medicine* 3.11 (2006), pp. 2011–2030.

[36] Marc Miravitlles, Patricia García-Sidro, Alonso Fernández-Nistal, María Jesús Buendía, María José Espinosa De Los Monteros, and Jesús Molina. "Course of COPD assessment test (CAT) and clinical COPD questionnaire (CCQ) scores during recovery from exacerbations of chronic obstructive pulmonary disease". In: *Health and Quality of Life Outcomes* 11 (2013), p. 147.

[37] Lucas Hermann Negri. *PeakUtils*. URL: `https://pypi.python.org/pypi/PeakUtils` (visited on 07/20/2016).

[38] Andrew L. Ries, Brian W. Carlin, Virginia Carrieri-Kohlman, Richard Casaburi, Bartolome R. Celli, Charles F. Emery, John E. Hodgkin, Donald A. Mahler, Barry Make, and Judah Skolnick. "Pulmonary rehabilitation: Joint ACCP/AACVPR evidence-based guidelines". In: *Chest* 112.5 (1997), pp. 1363–1396.

[39] John I Salisbury and Ying Sun. "Rapid screening test for sleep apnea using a non-linear and nonstationary signal processing technique". In: *Medical Engineering & Physics* 29 (2007), pp. 336–343.

[40] Scikit. *Generalized Linear Models*. URL: `http://scikit-learn.org/stable/modules/linear_model.html` (visited on 08/05/2016).

[41] Scikit. *Scikit-Neuralnetwork Documentation*. URL: `https://scikit-neuralnetwork.readthedocs.io/en/latest/` (visited on 08/14/2016).

[42] Heidi S Smith, Andrew J Criner, Dolores Fehrle, Carla L Grabianowski, Michael R Jacobs, and Gerard J Criner. "Use of a SmartPhone/Tablet-Based Bidirectional Telemedicine Disease Management Program Facilitates Early Detection and Treatment of COPD Exacerbation Symptoms". In: *Telemedicine and e-Health* 22.5 (2016), pp. 395–399.

[43] Juan Jose Soler-Cataluna, Miguel Angel Martinez-Garcia, Lourdes Sanchez Sanchez, Miguel Perpina Tordera, and Pilar Roman Sanchez. "Severe exacerbations and BODE index: Two independent risk factors for death in male COPD patients". In: *Respiratory Medicine* 103 (2009), pp. 692–699.

[44] Ioanna G Tsiligianni, Thys van der Molen, Despoina Moraitaki, Ilaine Lopez, Janwillem W H Kocks, Konstantinos Karagiannis, Nikolaos Siafakas, and Nikolaos Tzanakis. "Assessing health status in COPD. A head-to-head comparison between the COPD assessment test (CAT) and the clinical COPD questionnaire (CCQ)." In: *BMC pulmonary medicine* 12 (2012), p. 20.

[45] Jack V. Tu. "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes". In: *Journal of Clinical Epidemiology* 49.11 (1996), pp. 1225–1231.

[46] Peter Varady, Tamas Micsik, Sandor Benedek, and Zoltan Benyo. "A novel method for the detection of apnea and hypopnea events in respiration signals". In: *IEEE Transactions on Biomedical Engineering* 49.9 (2002), pp. 936–942.

[47] Mahesh Veezhinathan and Swaminathan Ramakrishnan. "Detection of obstructive respiratory abnormality using flow-volume spirometry and radial basis function neural networks". In: *Journal of Medical Systems* 31.6 (2007), pp. 461–465.

[48] Jadwiga A Wedzicha and Terence A R Seemungal. "COPD exacerbations: defining their cause and prevention". In: *Lanscet* 370 (2007), pp. 786–96.

[49] Qi Zhou. “A pulmonary rehabilitation system for COPD patients on mobile devices using the wearable RESpeck breathing and activity monitor”. MSc by research. University of Edinburgh, 2015.