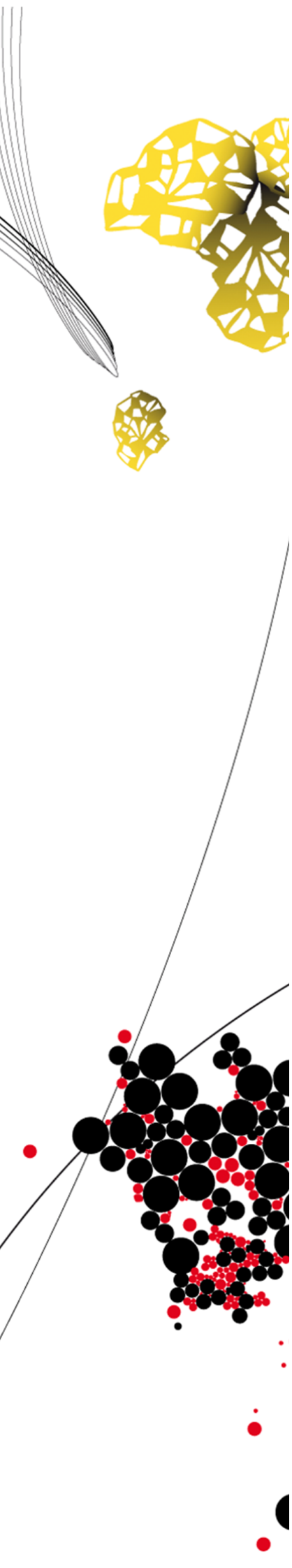


UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering, Mathematics & Computer Science

Automatic detection of user errors in spirometry data using machine learning techniques and the analysis of the effect of metaphors on the quality of spirometry measurements

Iris Heerlien

Final Thesis

Human Media Interaction

&

Data Science and Technology

07 2020

Supervisors:
dr. ir. R.W. Van Delden
dr. M. Poel

HMI and DMB group
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Preface

This thesis was conducted for the double master degree in Human Media Interaction and Data Science & Technology, which is a specialization of Computer Science. The thesis was carried out at the Human Media Interaction group of the University of Twente, in collaboration with PhD, Master, and Bachelor students in the fields of Technical Medicine and Creative Technology. Many people helped to bring this thesis to a successful end. First and foremost, I want to thank my supervisors, Dr. Ing. Robby van Delden and Dr. Mannes Poel, for inspiring me, and providing me with valuable feedback to bring my thesis to the next level. Furthermore, I want to thank PhD student Mattienne van der Kamp, master students Vivianne de With, and Arjen Pelgröm for explaining me the interesting world of spirometry. I also want to thank the participants and raters of the inter-annotation study for their effort, time, and enthusiasm.

Finally, I want to thank my family, friends, and housemates for their love and support, and for listening to my endless theories and ideas. Without them this thesis would not be possible.

Abstract

Asthma is the most frequently seen chronic disease among children [1]. Home monitoring of the asthma status of the children instead of at the hospital will give improved freedom. However, according to van der Kamp et al. (2017) [2], only 66% of spirometry attempts were performed technically correct at home compared to 92% when performed in the hospital.

The study described in this report is part of a broader project called Spiroplay, which goal is to improve the quality of home spirometry tests. This study focuses on three areas of this project. Firstly, an error detection algorithm based on machine learning techniques was designed and evaluated. To find the best model, different combinations of featuresets, hyperparameters, balancing techniques and machine learning techniques were evaluated. This process was repeated for three labelsets; a binary labelset consisting of two classes, one containing all technically correct attempts, and the other containing all attempts with errors, a combined labelset in which errorclasses are combined which are not directly linked to a criteria for a technically correct attempt stated by Miller et al. (2005) [3], and a third labelset in which the attempts preserve the label assigned to it. The results show that only the binary model, with a recall of 0.864 and a precision at 100% recall of 0.678, is useful in a real life system for the home monitoring of asthma.

The second area of focus is assessing the inter- and intra-rater agreement between professionals detecting errors in spirometry attempts. Three professionals labeled the same spirometry attempts. The inter- and intra-rater agreements were represented by the Cohen's Kappa score. The inter-rater agreement ranged from -0.123 to 0.380, which can be interpreted as a negative to minimal agreement. The intra-rater agreement was in the range of 0.648 to 0.860, which is a moderate to strong agreement. These results show that professionals detect different errors in spirometry data, showing that the rules on which the error detection is based are not strict enough. Therefore, before a generic error detection algorithm can be designed, the rules should be sharpened.

The third focus area of this research is the evaluation of the difference in quality of spirometry attempts when coached by a professional versus by a metaphor. The FVC, FEV_1, PEF values, and the number of errors were compared. No signifi-

cant differences were found between attempts coached by a metaphors and by a professional, however due to a possible research bias, the absence of a significant difference in the number of errors should be taken cautiously. These results imply that metaphors can be used to coach the children during home monitoring without significant quality loss based on PEF, FEV_1, FVC, and presumably the number of errors.

The conclusion of this research is that metaphors can be used as a coaching manner during home spirometry. However, before a generic error detection can be designed and used, the rules should be sharpened.

Contents

Glossary

ANN Artificial Neural Network.

$COPD$ Chronic Obstructive Pulmonary Disease.

$ECOC$ Error Correcting Output Codes.

EOT End of Test (seconds).

ERS/ATS European Respiratory Society/American Thoracic Society.

FEF Forced Expiratory Flow (liters/seconds).

$FEFV$ Forced Expiratory Flow Volume (liters).

FEV_1 Forced Expiratory Volume in the first second (liters).

FIF Forced Inspiration Flow (liters/seconds).

FVC Forced Vital Capacity (liters).

$FV\ curve$ Flow Volume curve.

$LSTM$ Long-Short Term Memory model.

MDV Mean Diurnal Variation (liters/seconds).

MSE Mean Squared Error.

$NAEPP$ National Asthma Education and Prevention Program.

PC_{20} The concentration of histamine or methacholine needed to result in a fall in FEV_1 of more than 20%.

PD_{20} The dose of histamine or methacholine needed to result in a fall in FEV_1 of more than 20%.

PEF Peak Expiratory Flow (liters/seconds).

RBF Radial Basis Function.

$RBFNN$ Radial Basis Function Neural Network.

SVM Support Vector Machine.

UAO Upper Airway Obstruction.

$VT\ curve$ Volume time curve.

Vbe Back Extrapolated Volume (liters).

Chapter 1

Introduction

Asthma is the most frequently seen chronic disease among children [1]. People suffering from asthma, also called respiratory disease, have difficulty breathing as their bronchial tubes narrow, swell, and produce more mucus than normally.

Different tests are available to diagnose or monitor asthma, such as spirometry, peak flow, methacholine challenge [1] , exhaled nitric oxide test, chest X-ray, CT scan, allergy tests, and sputum eosinophils [2] [4]. In this project, spirometry is used which provides physiological parameters of the flow and volume of air that is inhaled and exhaled. These parameters reveal episodic airway narrowing, which is the key feature of asthma. Besides, it is the most used objective monitoring tool of childhood asthma in hospitals at the moment [5]. Refer to section 2.2 for a more detailed explanation of spirometry.

At the moment, technologies are available to perform spirometry unsupervised. However, it was revealed by a home-monitoring study performed by van der Kamp et al. (2017) [2], that 66% of the spirometry measurements were performed technically correct at home compared to 92% when performed in the hospital. The goal of the project SpiroPlay, where this study is part of, is to improve the quality, expressed in technically correct attempts, of unsupervised spirometry, to support the professional in monitoring asthma patients. When the quality is as good as spirometry in the hospital, the spirometry tests can be performed at home which would give improved freedom to young asthma patients as they would have to visit the hospital less frequently to check the status of their asthma. Besides, it would reduce the cost of healthcare professionals as they can focus on the analyzing part of the spirometry tests instead of the complete monitoring process. The second goal of the project is

[1] During the methacholine challenge, the patient inhales increasing amounts of methacholine after subsequent tests. If the lung function drops by 20% or more before the maximum dose is reached, the patient is diagnosed with asthma.

[2] During sputum eosinophils, the saliva and mucus that comes out when the patient coughs is examined to identify high levels of white blood cells. If this level is high, the patient is diagnosed with asthma.

to acquire data about the patients more frequently, in order to monitor them more accurately.

The SpiroPlay project focuses on two areas to reach these goals. The first one is delivering feedback expressing if and which mistakes are made during a test by detecting the errors in the spirometry data instantly using rule-based artificial intelligence (explained in section 6.1.4), the second one is offering blowing metaphors to steer the behaviour of the patient based on the errors frequently made by the patient (explained in section 2.4).

The first focus area of the present study, which is part of the SpiroPlay project, is the error detection. Despite the hypothesized addition of the rule-based algorithm to the process, the present study goes one step further and targets on designing and evaluating a more holistic approach based on machine learning. The question 'How well can an error detection approach using machine learning techniques detect errors in spirometry data?' is answered.

The second focus area of this study is to evaluate the agreement in detecting errors in spirometry data by multiple professionals, answering the question 'What is the agreement in detecting errors in spirometry data by professionals?'. A low agreement points to a difference in detecting errors in spirometry attempts, complicating the designing of a generic error detection algorithm.

Thirdly, this study focuses on evaluating if coaching the children by metaphors instead of a professional during spirometry attempts result in a difference in quality. The question 'What is the difference in quality of the spirometry measurements between blowing behaviour coached by a professional and coached by a metaphor offered by the SpiroPlay system?' is answered.

If the error detection algorithm performs well, feedback can be given to the children based on the attempts during home spirometry, guiding the children to blow correct attempts. Besides, the quality of the measurements can be guaranteed by detecting whether an attempt was not technically correct and should be repeated. Additionally, if there is no quality loss in the spirometry tests when using the metaphors as coaching method, the metaphors can be used during home monitoring, instead of being dependent on the availability of a professional. Meeting one or both goals brings the medical world one step closer to home monitoring of asthma, resulting in lower healthcare costs and an improved freedom for the children.

The offering of metaphors in combination with an error detection approach come together in a system called SpiroPlay. An app is used in combination with a spirometer designed by NuvoAir (refer to section 2.3 for details). The app processes the data from the spirometer after an attempt is performed and determines instantly using the error detection algorithm if the attempt is an acceptable one, and if not, which errors are made. The errors made during attempts are counted and used to provide an

appropriate metaphor during a next attempt to support the person to overcome the frequently made errors. Additionally, after every attempt, the error made is shown (if an error is made) to help the child to focus on the part of the measurement the error occurred. The app also determines when three acceptable and two reproducible attempts are made and if the measurement is a successful one.

The remainder of this report starts with a chapter discussing background knowledge (chapter 2), followed by a literature review about unsupervised spirometry and related topics (chapter 3). Chapter 4 discusses related work about machine learning in monitoring and diagnosing asthma. The research questions of this study are stated in chapter 5, followed by the approach to answer these research questions in chapter 6. Chapter 7 shows the results when performing the approach, which are discussed in chapter 8. Chapter 9 answers the research questions and concludes the study.

Chapter 2

Background

This chapter discusses the disease asthma, the spirometry test, the spirometer, and the metaphors used in the SpiroPlay project.

2.1 Asthma

As mentioned, patients with asthma have difficulty breathing as their bronchial tubes narrow, swell, and produce more mucus than normally. This swelling due to inflammation results in extreme sensitivity to irritations. When irritated, the muscles around the airways tighten which might restrict the airways and trigger an overproduction of mucus [1].

The symptoms of people suffering from asthma differ per person. Symptoms are shortness of breath, pain in the chest, difficulty with sleeping due to coughing, shortness of breath, wheezing, and a whistling sound during exhaling. These symptoms can be triggered by e.g. allergens such as pollons or pet dander, strong odors, infections of the lungs such as flu, air pollution, tobacco smoke, exercise, changes in the weather, cold air, strong emotions, and medications. Also genetics play a role; if one parent has asthma, there is a 25% chance the child will have asthma as well. If both parents suffer from asthma, the chance is 50% [1, 6].

Asthma is seen in different strengths from mild intermittent asthma to severe persistent asthma. People suffering from mild intermittent asthma experience a few asthma attacks, symptoms during the night less than twice a month, and symptoms during the day less than twice a week while people suffering from severe persistent asthma have ongoing symptoms during the day and night, which are so frequent that it limits executing activities [7].

Next to a difference in severity, different types of asthma exist such as childhood asthma, adult-onset asthma, occupational asthma, exercise induced asthma, and seasonal asthma. In this research the focus is on childhood asthma. The

specific symptoms in childhood asthma include frequently less energy during play, shallow or rapid breathing, chest tightness, whistling sound during exhaling, retractions, shortness or loss of breath, tightened chest and neck muscles, tiredness, or weakness [8].

Asthma cannot be cured, however the symptoms can be controlled. As the symptoms vary over time, it is important to be monitored well [9].

2.2 Spirometry

Spirometry provides physiological parameters of the flow and volume of air that is inhaled and exhaled to reveal if the patient is suffering from episodic airway narrowing.

Several ways of performing a spirometry test exist, for example with or without an inhalation after exhaling. During the spirometry test in this study, the patient inhales deeply and exhales forceful and completely into a hose connected to a little device designed by NuvoAir (refer to section 2.3 for details). This test is called a Forced Vital Capacity (FVC) test. The device measures the flow in liters per millisecond during the exhalation. One way of presenting the data is a flow-volume curve. Figure 2.1 shows a curve of a child with asthma and a healthy child.

From the spirometry data, several values can be calculated. An example is the Forced Vital Capacity (FVC). Next to being the name of the test, it is the maximal volume of air one can exhale with maximally forced effort and from a maximal inspiration. Three values which are calculated from the FVC are the Forced Expiratory Flow (FEF) at 25% ($FEF_{25\%}$), at 50% ($FEF_{50\%}$), and at 75% ($FEF_{75\%}$) of the FVC. This can also be calculated of the inspiration phase; these are the Forced Inspiration Flow (FIF) at 25% ($FIF_{25\%}$), at 50% ($FIF_{50\%}$), and at 75% ($FIF_{75\%}$) of the FVC. Another value which can be calculated from the flow-volume curve is the Forced Expiratory Volume in the first second (FEV_1) which shows the maximum amount of air one can exhale in the first second of forced expiration after full inspiration. This value can also be calculated at other time points, such as the $FEV_{0.5}$, which is the Forced Expiratory Volume in the first half second, and the FEV_3, which is the Forced Expiratory Volume in the first three seconds. The ratio FEV_1/FVC is also a used value. Another value is the Peak Expiratory Flow (PEF or $PEFR$) which shows how fast and hard a subject exhales [3].

The measurements from a person are compared to comparable measurements from a healthy person with the same age, height, and ethnicity using the Global Lung function Initiative (GLI) table [11]. The values in this table are the ”predicted values”. By comparing the measured values collected during a spirometry test to the predicted values, one is able to calculate the ”percentage predicted values” which rep-

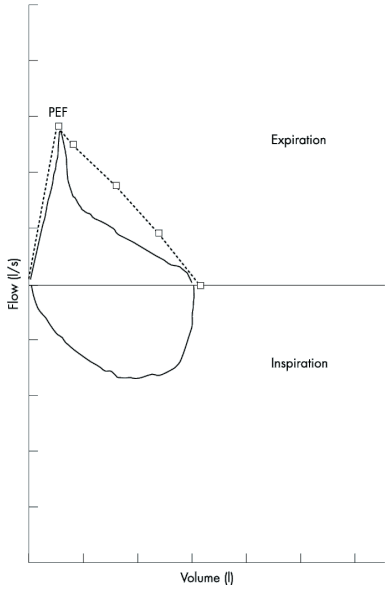


Figure 2.1: Flow volume curve of a child with asthma. The dotted line shows a curve of a healthy child. Source: Image 2 of Brand et al. (2003) [10]

resent how close the measurement is to a measurement from a comparable healthy person.

2.2.1 Criteria

A spirometry attempt has to meet a number of criteria to be technically correct. These criteria, based on Miller et al. (2005) [3], are described below and listed at the end of this section.

The test starts with a maximal inhalation. When this is not done maximally, the test is not acceptable as the lung capacity will not be measured correctly. To determine if the start of test is acceptable, the back extrapolation method is used. Using this method, the steepest slope on a volume-time curve is traced back in manual measurements (illustrated in figure 2.2), or using the largest slope averaged over an 80 ms period when using computerised measurements. The subject needs to have an extrapolated volume of less than 0.150 liter, or less than 5% of the expected FVC, depending on which one is greater.

When inhaled maximally, the patient has to exhale as forcefully as possible with minimal hesitation as hesitation reduces the PEF and FEV_1. After the burst of exhalation, the patient has to exhale maximally until the end of test (EOT). The end of test criteria are used to identify a good FVC effort. There are three EOT criteria: the first one is that the subject is not able to continue further exhalation. The second criteria is the volume-time curve showing no change in volume anymore ($<$ 0.025 L for $\geq$ one second). Lastly, the person should have exhaled for at least three seconds

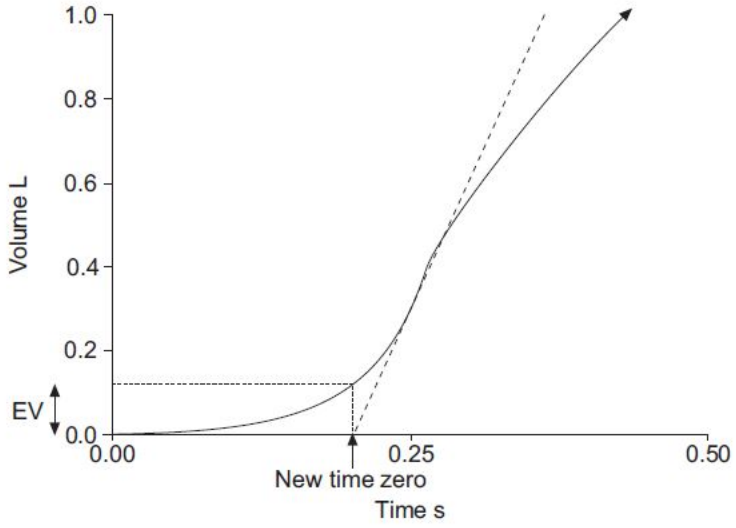


Figure 2.2: Back extrapolation by tracing back the steepest slope on a volume-time curve. The extrapolated volume is annotated in the image by EV. Source: Image 2 of Miller et al. (2005) [3]

when the subject is aged below ten years, and six seconds when aged above ten years. When these criteria are not met, the FVC value cannot be used. However, the FEV_1 can still be used in some situations as this is only about the first second of the measurement [3].

Additionally, some overall criteria apply [3]. First, coughing will make a test unacceptable, as well as glottis closure, an extra breath taken during the attempt, or hesitation during the attempt which causes a stop in airflow in a way that it influences the measurements. Besides, the lips should be sealed around the mouthpiece to prevent leak, and the tongue and teeth should not occlude the mouthpiece.

All criteria named above are within-manoeuvre criteria. Next to these criteria are between-manoeuvre criteria; an adequate test should consist of a minimal of three acceptable FVC manoeuvres and two reproducible ones. Two attempts are reproducible when the difference between the FVC values met at two manoeuvres is smaller than 0.150 liter. This should also hold for the FEV_1 values. If the subject has a measured FVC of lower than 1.0 liter, the difference in FVC and FEV_1 values should be lower than 0.100 liter. In case these criteria are not met in the first three attempts, the subject can perform a maximum of five more attempts according to Miller et al. (2005) [3].

In conclusion, the list with criteria for a technically correct attempt is as follows:

1. Maximal effort

2. Maximal inhalation

3. Minimal hesitation at the start

4. Duration of attempt is $>$ 3 s, no plateau in the VT curve, and the person should not be able to continue exhaling

5. No cough during the attempt

6. No glottis closure

7. No extra breath taken during the attempt

8. No hesitation during the manoeuvre

9. No leak

10. No obstructed mouthpiece

11. Three of the attempts are acceptable.

12. Two of the attempts are reproducible

2.3 Spirometer

The Air Next spirometer (see figure 2.3) of the company NuvoAir is used in the SpiroPlay system. NuvoAir is a digital health start-up focused on respiratory care. The small and portable spirometer measures respiratory flow and can be connected to smartphones/tablets via a Bluetooth connection. The respiratory flow is measured by setting a rotor in motion by exhalation in the turbine connected to the spirometer. Infrared interruption is used to determine the airflow rate and volume. The flow range is between zero and fourteen liter per second. The flow and volume can be determined with an accuracy of respectively approximately 5% and 3%. Spirometry parameters, such as the FEV_1, FVC, the FEV_1/FVC ratio, and the PEF, are calculated from this flow and volume data and presented in the corresponding app.

NuvoAir is ISO 13485:2016 certified and the Air Next spirometer is CE certified as a class IIa medical device [12].

Figure 2.3: The Air Next spirometer of NuvoAir [12]

2.4 Metaphors

As explained in chapter 1, the app will offer blowing metaphors to encourage the patient during the attempt, and to steer his or her behaviour. The metaphors used in the studies of this research are shown in figure 2.4. When using the metaphor of the car, the car starts, the pointer of the tachometer moves from left to right, and the vertical bar fills up in green during inhalation. During exhalation the car drives and changes from a normal car into a sports car. If the predicted FVC is met, the car will pass the finish. The second metaphor shows a springboard diver who jumps during inhalation, and do tricks when going down during exhalation. The third metaphor presents a bow and arrow. During inhalation, the arch is stretched and the vertical bar fills up in green, during exhalation the arrow punctures the balloons. The better the attempt, the more balloons will be punctured.

(a) The car *(b)* The springboard diver *(c)* The bow and arrow

Figure 2.4: The metaphors used during the studies of which the data is used in this research.

When the system will be used in home monitoring, the patients will make use of several metaphors during the spirometry tests. More metaphors will be provided over time to keep the children engaged. The metaphors used will be tailored towards the child based upon the error made frequently by this child. For example, if a child has difficulty breathing out long enough, a metaphor which focuses on this part of the spirometry test will be offered.

Chapter 3

Literature review

This chapter describes previous research in unsupervised spirometry and related topics. The systematic approach used to find information about these topics is elaborated in section 3.1. The answers to the questions asked in this section can be found in section 3.6.

3.1 Method literature review

The first step in the literature review was to find information about spirometry in children. The questions to be answered were:

1. What are the differences between spirometry in adults and children?
2. What do the differences between spirometry in adults and children imply?
3. How are the consequences of the differences between spirometry in adults and children dealt with?

The creator of the rule-based error detection approach (explained in section 6.1.4), V. De With, recommended the following papers in the area of spirometry in children: Tomalak et al. (2008) [13], Loeb et al. (2008) [14], Miller et al. (2005) [3], and Thompson et al. (2006) [15]. The papers by Tomalak et al. (2008) [13] and Loeb et al. (2008) were used in section 3.2 to answer the questions above.

The second subject discussed is home spirometry:

1. What is the quality of the spirometry data derived during home spirometry?
2. What procedure related aspects influence the quality of home spirometry?

The search term used was ”spirometry at home” and the search engine ”Google Scholar” was used. The results with ”asthma” and ”home spirometry” or alike (e.g.

self-recorded, self-management, home monitoring, portable) in the title or abstract were evaluated. Using these inclusion criteria, 5 papers were selected which were read and summarized. Interesting sources used in these papers were evaluated and, if useful, summarized as well. This process was repeated two times. As interesting information was found about the quality of PEF, information was searched about the quality of other values which can be calculated from spirometry data such as FEV_1, FVC, and $FEF_{25\%-75\%}$. This resulted in a third question:

Which values derived from spirometry data are useful in monitoring or diagnosing asthma?

The search terms used for this were:

1. correlation FEV_1 FVC asthma
2. parameters in asthmatic children FEV_1 FVC
3. objective parameters asthmatic children
4. FEV_1 FVC "more sensitive test"
5. "clinical features" children with asthma
6. "FEV_1 FVC" clinical and physiologic parameters

The process of reading papers, and finding interesting sources used in these papers was again repeated two times. After this, the questions were answered elaborately.

The third subject of this literature review was "spirometry and games" as metaphors are used in the product to help the children to overcome their errors. The question to be answered is:

1. How do games used during a spirometry attempt influence the quality of the spirometry data?

The papers found when looking for information about "spirometry in children" mentioned games as well. The sources about gaming found in these papers were used to start the literature review of "spirometry and games". From these papers, sources were used to find new papers. This process repeated itself two times after which the question was answered.

The last subject of this literature review is "Inter- and intra-rater agreement when assessing spirometry data". To be able to interpret the results of the to be designed

error-detection algorithm, the agreement between two observers, and one observer over time, have to be known. The questions to be answered were:

1. What is the agreement between two professionals when assessing (the errors in) spirometry data?
2. What is the agreement of one professional over time when assessing (the errors in) spriometry data?

The search terms used are shown in the enumeration below. The search terms were also used when looking for papers about intra-rater agreement. 'Inter-rater' was then replaced by 'intra-rater'.

1. Error detection spirometry inter-rater agreement
2. Error detection spirometry inter-rater reliability
3. Error detection spirometry inter-observer agreement
4. Error detection spirometry inter-observer reliability
5. Error detection spirometry inter-rater response agreement

Around fifty useful papers were found in total. These papers gave insights in the area of unsupervised spirometry, also in combination with games. This revealed challenges showing that the proposed approach of this project contribute to existing research.

3.2 Spirometry in children

Most spirometry test use the European Respiratory Society/American Thoracic Society (ERS/ATS) criteria [16] for determining the acceptability of spirometric measurements. However, the questions are if these criteria are different for children, what these differences imply, and how the consequences are dealt with. The researches discussed in this section answer these questions.

Tomalak et al. (2008) [13] studied if criteria for adults were met by children below ten years. 233 children were tested of which 116 children (all but one under seven) did not finish the experiment. Reasons were not understanding the procedure, a lack of peak expiratory flow in the beginning of the test, variable or sub-maximal respiratory efforts, and lack of interest. The tests were performed by experienced personnel. It was found that the Vbe criteria (150 ml and 5% of the FVC at the start of the attempt) was met bij 80.4% of the children. There was a weak, however significant relation between age and Vbe. The second acceptability criteria is ambiguous:

the time to PEF should be "short". Seventy-two children (61.5%) had a time to PEF of less than 100 ms which is stated to be acceptable. This did not correlate with age. The FET values, which is acceptable when bigger than three seconds, were in the range of 0.71 to 6.9 seconds. However, only 23.9% of the children had a FET bigger than three seconds. In twelve children between 5.3 and 8.5 years, the FET was even below one second. Besides, it was found to be significantly correlated with age. The FEV_1 reproducibility criteria (a difference in FVC and FEV_1 of less than 150 ml, or 100 ml if FVC and FEV_1 are below 1000 ml, when comparing the two best measurements) was met by 101 children and was negatively correlated with age. The FVC criteria was met by 105 children and was not correlated with age. Both reproducibility criteria were met by ninety-two children (78.6%). When all four criteria (Vbe less than 150 ml and 5% of FVC, a "short" time to PEF, FET bigger than three seconds, and a difference between the two best values of FVC and FEV_1 of less than 150 ml, or 100 ml when the FVC or the FEV_1 is below 1000 ml) are combined, 17.1% of the children met the criteria. If the criteria for FET was left out, this percentage is 63.2%.

In the research by Loeb et al. (2008) [14], 393 children in the age of four to seventeen years old are asked to perform spirometry for the first time. The tests were performed under supervision of one or two respiratory therapists. A maximum of eight attempts to reach an acceptable test was allowed. The criteria used were specific criteria for children of six years old or younger, based upon Miller et al. (2005) [3], and Beydon et al. (2007) [17]. These revised criteria include that a start of test is acceptable if the extrapolated volume is less than 80 ml, or 12.5% of the FVC. If this is not met in preschool children, this is not directly an indication to exclude the attempt. Besides, the plateau in the end of test criteria is not defined for preschool children. However, the flow-volume curve needs to demonstrate a fast rise to peak flow in combination with a smooth descending limb. When looking at between-manoeuvre criteria, the preschool children only need to have two acceptable tests, and the difference between the FVC and FEV_1 needs to be within 100 ml, or 10% of the highest value. 292 children (74%) met the revised ERS/ATS criteria for an reproducible and acceptable test. This increased with age and was above 50% by the age of six and reached a plateau of approximately 85% at the age of 10. The success rate was not influenced by the gender or race of the children. In preschool children (four to six years old), the criteria which caused the most unacceptable tests were glottis closure and non-maximal efforts (38% of mistakes made each), and premature termination (19% of mistakes made). In school-aged children (seven-seventeen years old), these criteria were failure to plateau (49% of mistakes made), premature termination (17% of mistakes made), glottis closure, or non-maximal effort (13% of mistakes made each). If the guidelines for school

age children was used for six years old, the success rate would only be 42%. The researchers conclude that it is possible for children to perform acceptable and reproducible spirometry on their first effort when using the revised ERS/ATS criteria for preschool children.

Conclusion

From these studies, we can conclude that it is necessary to use specific criteria for assessing the spirometry attempts by children. Examples are changing the criteria of an acceptable FEV_1 to 80 ml, or 12.5% of the FVC, and accepting a difference of 100 ml, or 10% of the highest value between the FVC and the FEV_1 when looking at between-manoeuvre criteria. Also, a less strict criteria than requiring a FET of three or more seconds should be used as this criteria was only met by 23.9% of the children in the research of Tomalak et al. (2008) [13]. They suggest to revise the criteria.

3.3 Spirometry at home

Above research is performed under supervision of experts. In this project, a portable spirometer is used at home, which means no supervision, no quality check, and no encouragement by a professional. Several challenges appear in this situation, creating questions such as: what is the quality of home spirometry, what procedure related factors influence this quality, and which values derived from spirometry data are useful when monitoring or diagnosing asthma? Researches discussing these topics are discussed in this section.

3.3.1 Quality of home spirometry

Performing a good spirometry test can be very hard; one has to inhale deeply and exhale forcefully and long. As it is so hard, a lot of measurements contain errors. This section focuses on the quality of home spirometry measurements, the errors made, and the comparison between the quality of home spirometry and in-office spirometry.

The quality of spirometric data has to conform international guidelines [16]. Reddel et al. (1998) [18] assessed if self-recorded spirometric data met these guidelines. Thirty-three subjects between 18.6 and 67.4 years old were asked to perform spirometry measurements twice daily, the morning one immediately after waking up. The subjects were instructed how to use the spirometry device before the experiment. The within-session reproducibility of FEV_1, FVC, and PEF was calculated

during the first and 9th week of budesonide[1] treatment. An excellent reproducibility was found; 90% of the sessions met the reproducibility criteria when looking at FEV_1. However, they also found that it is difficult to control the quality and state that an accompanying paper diary is still necessary. This paper diary should be used to write down e.g. symptoms or factors (e.g. severe facial pain) that may have influenced the measurements.

Gannon (1999) [20] compared supervised and unsupervised PEF recordings. forty-four participants in the age range of fifteen to sixty-five years were trained after which they were asked to record their PEF every two hours during waking hours for two to three weeks between two clinic visits. During the clinic visits, they performed two unsupervised measurements, one supervised, and one supervised measurement during which they were encouraged to e.g. exhale maximally. When comparing the unsupervised measurements to the supervised measurements with encouragement, a decrement of twenty-one liter per minute was found when recording unsupervised. When comparing the supervised measurement with encouragement to the supervised measurement without encouragement, a decrement of another nine liter per minute was found. Also, a deterioration was seen in 54% of the PEF measurements. According to the authors, these detoriations could be due to a lack of effort or technical reasons.

In research executed by Thompson et al. (2006) [15], self-administered spirometry is performed at home using the hand-held device "ndd EasyOne Frontline Spirometer". This device saved all data, measured electronically the quality of the manoeuvre by detecting when the acceptability and reproducibility criteria were met, and showed on-screen instructions based on the criteria not met during the last attempt. The participants were trained how to perform spirometry manoeuvres in home for five days, one hour on the first day, and fifteen minutes to half an hour on the last four days. The ATS criteria were used with some amendments as it is found that the criteria for adults cannot be applied directly to children [13, 14, 21, 22]; the FET was lowered from six to four seconds, and the end of test criteria used was the end-of-test volume (EOTV). The end of the test was marked when the inspiration was more than 150 ml or the volume change was less than 45 ml over two seconds if FET was lower than four seconds, or 60 ml in case FET was more than four seconds. Besides, the FEV_1 and FVC repeatability was set to 10% instead of 5% and PEF at 20% instead of 10%. Next to using criteria, the curves were evaluated visually. The participants (sixty-seven children between nine and eighteen years old) were asked to perform the measurements in the morning, afternoon, and evening, and to complete a diary every two waking hours. A maximum of six attempts per mea-

[1] Budesonide is a medication of which the most important substance is corticosteroid. This medication prevents swelling in the lungs which decrease the severity of an asthma attack. [19]

surement was allowed. Besides, two groups were compared of which one had daily follow-ups for ten days, and the second group weekly for two months. The overall quality was always higher than 75% when evaluating the manoeuvres based on the three flow-volume criteria. The most common mistakes in this age group were abrupt ending (0.93% of the total manoeuvres) and invalid time to peak expiratory flow (PEF) (1.03% of total manoeuvres). The most common mistakes in the visually rejected manoeuvres were variable effort (6% of the total manoeuvres when having daily follow ups, and 3.93% when having weekly follow ups), often in combination with glottis closure (0.7% of total manoeuvres when having daily follow ups, and 1.2% when having weekly follow ups) and cough (1.0% of total manoeuvres when having daily follow ups, and 0.8% when having weekly follow ups). They found that compliance was not significantly higher for the group with daily follow ups (more than 90% vs. more than 84%) showing that doing spirometry at home is a good option. They also found that the quality of the manoeuvres was significantly lower for nine to twelve aged children compared to a group of thirteen to eighteen aged children. Additionally, the paper indicates that quality assurance was increased by showing correcting instructions for the next manoeuvre based on the errors made during earlier attempts.

Mortimer (2003) [23] compared a portable spirometer and an office based spirometer to evaluate if the portable spirometer gave reliable results. The two spirometers were validated in an office after which a two-week home study was performed using the portable spirometer. The ninety-two participants were between six and eleven years old. The portable spirometer also included a program to help the children during their measurements by showing in text why a measurement was not acceptable. They found that the overall agreement between the software/portable spirometer and the physician/office spirometer was 74%. The office based spirometer counted that 43% of the sessions had at least three acceptable curves and at least two reproducible curves. According to the software of the portable spirometer, 67% of the sessions had three acceptable and two reproducible curves. FEV_1 and PEF had the best agreement; there was not any systematic difference found between the two devices. The difference in agreement for FVC was due to small differences in the implemented algorithms or to the difference in sensitivity of the devices at very low flow rates. During the two week home study, 59% of the sessions produced three acceptable curves and two reproducible ones. This was significantly higher (65% instead of 47%) when the participants were eight years old or above. These results are comparable to the results met in the office sessions. 25% of the curves which were accepted by the portable device were rejected by the physician as the software was not able to find problems in the mid-portion of the curve. It was concluded that although the agreement was high between the two devices used in the office, the

software should be programmed so that it is able to find problems in the mid-portion of the curve.

Conclusion

In conclusion, spirometry at home is possible, however the mid-portion of the curve should also be examined for quality [23] and a diary is still necessary [18]. The main error found by Gannon (1999) [20], with a target group of patients between fifteen and sixty-five years old, was a lack of effort. The main errors found by Thompson et al. (2006) [15], having a target group of children between nine and eighteen years old, were abrupt ending, invalid time to PEF, variable effort in combination with cough and glottis closure.

Other conclusions are that encouragement improves the quality of the measurement [20], and that showing correcting instructions after a non-acceptable attempt increased the quality assurance of the next attempts [15].

3.3.2 Compliance

Asthmatic people can monitor their asthma at home in different ways. One is keeping a diary in combination with performing a spirometry test. However, the date and time of the measurements is not saved in several monitoring situations making it impossible to check whether the measurements are performed on time. Chowienczyk et al. (1994) [24] shows via an experiment with thirty-three adults between nineteen and seventy-eight years old that when people are not aware of the fact that the date and time is stored, measurements are taken at wrong moments in time, invented, or taken all at once, just to complete their diary. This research showed that if people know that their data from the spirometers is electronically recorded, they will perform the measurements on time more often.

Wensley et al. (2001) [25] assessed the compliance during unsupervised spirometry, and the quality of spirometric data taken unsupervised at home. Ninety asthmatic children in the age of seven to fourteen years old took part in this study. They were asked to perform spirometry tests twice daily for a period of sixteen weeks. They were taught how to perform a spirometry test on the first day of the test. Every four weeks, the patients were visited to download the data from the spirometers and to retrain the patients if needed. FVC, FEV_1, PEF and $FEF_{25\%}$ and $FEF_{25\%-75\%}$ values were assessed. Only the expiratory manoeuvres were collected. The ATS criteria were used for assessment of the measurements. What was found is that the children became less compliant month by month; 81.4% completed the tests during the first month, 78.4% during the second month, 71.4% during the third month, and

70.3% during the fourth month. This decline in compliance resulted in a decrease of valid data over time. The technical quality of the data stayed the same, however there were big individual differences between the children. These results show that the decline in valid data was due to compliance instead of loss of skill. They conclude that spirometry at home is possible, but not for a long period of time. According to their results, a period of 4 weeks is optimal.

Other results are found in research done by Pelkonen et al. (2000) [26]. They evaluated the reproducibility of spirometry measurements taken at home by a group of children (110 participants between five to ten years old) who were newly diagnosed with asthma. The measurements were assessed based on the ATS criteria. These criteria were not revised to criteria found to be more suitable for children. The children performed spirometry tests twice daily for twenty-four days, logging the FVC, FEV_1, and PEF score combined with time and date of the measurement. It is unclear if they were aware of the logging of the date and time. A compliance of 94% was found. 77% of the measurements were reproducible. However, a big individual variation was found in the range of 21 - 100%. When splitting the group in smaller groups based on age, it was seen that the reproduciblity increased with age; the five-six year age group had a mean spirometry reproducibility of 72.8%, while the age group seven-eight years old had a mean score of 77.1% and nine-ten years old had a mean score of 84.5%. They conclude that home spirometry is possible, however also 23% of the measurements were not acceptable or reproducible which still is a concern. The compliance and reproducibility did not change over time. The difference in results between Wensley et al. (2001) [25] and Pelkonen et al. (2000) [26] can be due to novelty [25]. Besides, the research of Pelkonen et al. (2000) [26] only lasted twenty-four days which makes it hard to compare the two studies as the study by Wensley et al. (2001) [25] only measured compliance per four weeks. Besides, it is unclear if the participants from the study by Pelkonen et al. (2000) [26] knew the date and time of their measurements were saved.

Redline et al. (1996) [27] analyzed if children, age five to nine years, from inner city areas in America were able to initiate and maintain peak flow recordings in a paper diary for three weeks. They were given a recording meter of which one was a covert meter of which the children and parents was not told that the data was saved automatically. The missing values in the paper diary were compared to the missing values obtained from the meter. It was found that the number of missing entries in the diaries increased from 1.4% to 10.6% comparing the first and third week and that the meter showed a significantly greater percentage of missing data than the paper diaries which also increased over time. In the third week, 52.4% of the records were missing from the meter, and 15.3% from the paper diaries. Also, some values were not transcribed right. This shows that the children and caretakers from this

subgroup of society have difficulty in maintaining these peak flow recordings and that the manual records are not always reliable. This increases over time. The authors suggest to shorten the period of home monitoring to two weeks as this may help to increase compliance. Also, the children and caretakers technique in recording PEFs should be monitored. Another reason given for the decrease in compliance is that the participants did not have much opportunity to develop rapport with the personnel of the study, and the fact that the study did not require a lot of commitment. It could also be that the participants were too young. Another reason given is that the children were told to be compensated financially regardless of how well they completed their PEF diaries.

Another research which focuses on the comparison between electronic and paper diaries was performed by Hyland et al. (1993) [28]. Both diaries were completed twice daily at home for fourteen days. The electronic diary asks, next to measuring PEF, to fill in questions about symptoms. The participants were not told that the electronic data was stored. It was found that thirty-two retrospective entries were made and that 15% of the written values were not the same as the measured PEF values; 75% of the participants had at least one discrepant entry. PEF variations was related significantly to the number of missing days and the number of discrepancies. Around 20% of the written entries had errors. The conclusion the authors made was that the reason for the poor diary completion could come from the unreasonable expectations the doctors have of patients, and that incomplete instructions were given. They mentioned that electronic diaries could result in better quality of records in combination with instructions what to do when a day is missed, and a feature which accommodates the forgetfulness of people.

Verschelden et al. (1996) [29] analyzed the compliance and accuracy of home PEF measurements. Twenty adults were asked to measure PEF twice a day. The used device stored the PEF data automatically. The participants were not aware of this. The duration of the experiment was forty-four to 131 days. It was found that 54% of the to be measured values were written down, and 44% were really measured; 10% of the to be measured values were not according the written down values. The best compliance was during the first two weeks after which it decreased and the number of invented values increased. The compliance decreased to 40% after one month, and reached a plateau of 35% shortly after that. The conclusion of this research was that the compliance with PEF measurements is poor in stable asthmatic subjects over a three month period, and that 22% of the values that are written down is invented. Some solutions were given such as reinforcing the need of PEF monitoring, give instruction on a treatment plan based on the measured values, or ask the participants to measure PEF only when their symptoms increase.

Conclusion

These researches show that home spirometry is possible, ideally for a period of four [25] or two weeks [27]. Afterwards the compliance is likely to drop [25]. Additionally, the researches reveal that paper diaries can be unreliable as there is no check whether the patient took the measurements at the right time. Besides, data is invented by the patients [24, 27, 29]. Also, Pelkonen et al. (2000) [26] showed that compliance could be increased by a novelty effect.

3.3.3 Usefulness of measured values

Several values can be extracted from spirometry data, such as PEF, FEV_1, and FVC. However, do these values really say something about asthma severity, and do they add knowledge next to monitoring symptoms? This section reviews research in this field.

PEF

The Peak Experiratory Flow (PEF) is used often when monitoring asthma. However, the usefulness of this value is questionable. This subsection discusses several researches performed using PEF.

Brouwer et al. (2006) [30] examined if the peak flow and FEV_1 score relates to other estimates of asthma severity in children. An electronic home spirometer was used which stored the data automatically. thirty-six children in the age group of six to sixteen years completed this research. They all knew beforehand how to perform spirometry. Before the twice daily spirometry measurements, they were asked to record their asthma severity score on a scale. The FEV_1 was expressed as a percentage of the predicted value, the asthma severity score and PEF as a percentage of the personal best value. The variation in FEV_1 and PEF was communicated in terms of the size of the day's range as a percentage of the day's mean. The results show that PEF and FEV_1 measurements of this research did not correlate significantly to the asthma severity score or the patient's quality of life score. It was even the case that increases in the severity score correlated with decreases in PEF and FEV_1 scores for some patients, but by increasing values in others. They also found that the concordance between PEF and FEV_1 is low; only 67% showed an acceptable concordance. Therefore, they conclude that electronically recorded scores are not clinically useful as they are too inconsistent with other asthma parameters. One reason given for these poor correlations is the lack of quality control of the measurements at home, however earlier in section 3.3.2, we found that home spirometry recordings in children are most of the time acceptable [25].

Sly and Flack (2001) [31] found frequent discrepancies between true lung function and PEF measurements; only six from fifteen clinically important deteriorations in lung functions were found. The discrepancies went both ways; a fall in PEF did not always mean a fall in lung function, and a fall in lung function was not always shown as a fall in PEF. It is mentioned that FEV_1 could be a better measure of lung function when home monitoring. However, the statement is made that next to the accuracy of the value measured, there are other problems such as compliance and technical expertise in performing the measurements.

Brand et al. (1997) [32] performed research to find out if there are relations between PEF and symptoms, airways hyperresponsiveness, level of lung function, and atopy. 116 asthmatic children in the age range of seven to fourteen years old were asked to record their symptoms and PEF twice a day for a period of two weeks.They were all checked afterwards if they used the right technique, and 102 did so, and also completed the diary. From the results of these 102 children, it was found that atopy and FEV_1 were not significantly related to variation in PEF. However, PD_{20}, the dose of histamine needed to result in a fall in FEV_1 of more than 20%, and symptoms were weakly, however significant, related to PEF variation. This shows that none of the values on its own gives a complete overview of the lung function of a patient.

Another study by Brand et al. (1999) [33] looked into the relation between PEF variation of a patient over time and the percentage of days without symptoms, FEV_1, and PD_{20}. The FEV_1 and PD_{20} were measured bimonthly, PEF and the symptoms scores twice daily during a long term treatment using inhaled corticosteroids in 102 children age range seven to fourteen years. It was found that PEF variation had a poor concordance with the other parameters. It can be concluded that only monitoring PEF may be insufficient to measure asthma severity in children and clinically relevant deteriorations in other parameters may be missed.

Another research done in this area is performed by Gern et al. (1994) [34]. The PEF variation, symptoms, methacholine reactivity, and medication requirements were compared in seventy-four children in the age range of five to twelve years old to look for a relationship between phenomena. A significant correlation was found between Mean Diurnal Variation[2] (MDV), which is a way to calculate PEF variation, and symptoms, and between MDV and methacholine reactivity. They concluded that the correlation between PEF variation and other variables is statistically significant, however these relations are too weak to be useful in the treatment of the patients. They also state that MDV could be a useful indicator of asthma severity.

Ferguson (1988) [35] compared symptom score and PEF readings to FEV_1 and

[2]MDV is calculated by $\frac{PM-AM}{(PM+AM)/2} * 100$, where PM is the evening measurement and AM the morning measurement

$FEF_{25\%-75\%}$ values. The two latter values were measured every two weeks during sixteen weeks in twenty children in the age group of six to fourteen years old, the symptoms and PEF readings were written down twice a day. The symptoms score was calculated by scoring the severity from zero (no symptoms) to four (wheeze, cough, and dyspnea requiring hospitalization) and adding the frequency of episodes of symptoms. This frequency was represented by a score between zero to eight showing the durance of an episode (zero meant no attacks, eight meant episodes longer than six hours). The results showed that the symptoms scores were significantly associated with a decrease in low peak flow days and mean PEF, however not with a decrease in FEV_1 and $FEF_{25\%-75\%}$. They found that PEF readings are useful next to symptom diaries as symptoms are subjective while PEF readings are objective measures. However, the values on itself are not adequate for assessing the variable airway obstruction. They state that PEF readings may provide helpful information if recorded twice a day, however this asks for excellent cooperation from the child which can be difficult at home. They state that, although it did not significantly correlate with symptoms, $FEF_{25\%-75\%}$ was a more sensitive indicator of airflow obstruction in comparison to PEF, FEV_1, and symptoms. A decrease in $FEF_{25\%-75\%}$ could be measured when there were no symptoms or a change in peak flow rates. In combination with the fact that there is a high probability of persisting airway obstruction even when there are no symptoms, and normal peak flow rates, a change in $FEF_{25\%-75\%}$ gives valuable information. One of the reasons that it changes when PEF does not is that $FEF_{25\%-75\%}$ score is almost independent of the effort. Another reason which is given is that different from FEV_1 and PEF, which are measures of airflow in the central airway, $FEF_{25\%-75\%}$ is a measure of airflow in the peripheral airways.

Self-management of PEF

Measuring PEF values (on a daily basis) is used in treatment procedures to monitor the asthma severity of patients. However, it was demonstrated by several studies that using measurements from peak flow meters did not improve asthma outcomes compared to people who are taught to use their symptoms to self-manage their asthma [36], or people who receive conventional treatment [37].

Wensley and Silverman (2004) [38] performed research to find out if knowledge of PEF enhances self-management of asthma. Ninety children in the age range of seven to fourteen years were divided into two groups; one group which received symptom-based management, and one group which received management based on symptoms and PEF. The latter group was asked to perform spirometry twice a day for a period of twelve weeks and to keep a symptom diary once a day. No

differences were found in symptom scores, QoL, lung function, or health service between the children in the two groups. It was found that the children responded to changes in their symptoms to change their medication, instead of to changes in PEF.

Another research in this area is performed by Agertoft and Pedersen (2000) [39]. They found that when asthma is self-managed by adults, the health outcomes were comparable when either PEF or symptom scores were used. The factors that improved the health of the participants were education in self-management including a written action plan, regular medical review, and self-monitoring using either PEF or symptoms.

In another study focusing on children it was found that PEF monitoring did not have additional benefit over the daily recording of symptoms, and the used of bronchodilators [40].

Other values

As touched upon earlier, other values such as FEV_1, FVC, and $FEF_{25\%-75\%}$, and ratios thereof, can be useful to calculate from spirometry data next to PEF. The usefulness of these values is reviewed more extensively in this section.

In research performed by Ramsey (2005) [41], the relationship between several spirometric measures and asthma severity was examined. 438 children in the age range of four to eight-teen years were included in the research. Their asthma severity was based on a questionnaire which was in accordance with the National Asthma Education and Prevention Program ($NAEPP$) guidelines. The predicted values of the ethnic-specific NHANES 3 [42] were used, except for the Puerto Rican children. For these children the predicted values for Mexican Americans were used as there were no values available for Puerto Ricans. They found that the FEV_1/FVC ratio decreased significantly in children with severe asthma versus children with mild asthma. The FEV_1 percentage predicted value was significantly lower in children with severe asthma, and the FVC was significantly higher in patients with severe asthma in comparison with patients with mild asthma. However, this difference vanished when FVC percentage predicted value was used instead of FVC. Furthermore, after adjustments to amongst others individual allergens and race, it was found that only the FEV_1/FVC ratio is a useful indicator of the asthma severity in children.

The association between FEV_1 percentage predicted value and the risk of an asthma attack in the year after a taken spirometry test is examined by Fuhlbrigge et al. (2001) [43]. 13,842 children, fifteen years old at maximum, were tested every year for a period of fifteen years. Until an age of fourteen, the parents filled in

the questionnaires. When the children reached this age, they were allowed to fill in the questionnaires themselves. A strong association was found between FEV_1 percentage predicted value and asthma attacks in the year after the taken test; an increase in FEV_1 percent predicted led to a decrease in asthma attacks. From the group where the parents filled in the questionnaire, 60.4% of the children with an FEV_1 percent predicted score below sixty reported an attack while just 25.4% of the children having an FEV_1 percent predicted score above 80% had an attack. A similar relationship was seen when the children themselves filled in the questionnaire; 73.9% of the children reported an attack while having an FEV_1 percentage predicted value below sixty, and 29.4% when this score was above 80%.

In an essay written by Spahn et al. (2004) [44] the question is asked if FEV_1 is the best measure of asthma severity in childhood asthma. The answer is clearly no. Children with normal FEV_1 values can still have asthma. This is more a rule than an exception. The reason is that asthma is a slowly progressive disease; it is found in adults that a decline of ca. 1% of predicted FEV_1 per year is seen. As children are young and thus have asthma for a relative short period of time, their FEV_1 can still be normal during periods of stability. Asthma diagnosis and treatment should not be solely based upon the FEV_1 value as then children will falsely be diagnosed with not having asthma, or will be undertreated.

Other studies focus on the value of the $FEF_{25\%-75\%}$ measurement. The goal of Simon et al. (2010) [45] was to determine if the $FEF_{25\%-75\%}$ percentage predicted values offers advantages over FEV_1 percentage predicted values, or over the ratio FEV_1/FVC percentage predicted values in the evaluation of childhood asthma. $FEF_{25\%-75\%}$ is less sensitive for effort and thus may give more stable results. Besides, $FEF_{25\%-75\%}$ is a measure of the airflow in the peripheral airways instead of the central airway as the FEV_1 and PEF are. Data from the Pediatric Asthma Controller Trial, and the Characterizing the Response to a Leukotriene Recepter Antagonist and Inhaled Corticosteroid trials was used. Data from 437 children between six to seventeen years old were included in this research. They found that the $FEF_{25\%-75\%}$ percentage predicted values and the FEV_1/FVC percentage predicted values were positively correlated with morning and evening PEF percentage predicted values, and negatively correlated with log_{10} fraction of exhaled nictric oxide and bronchidilator responsiveness. The $FEF_{25\%-75\%}$ percentage predicted values and the FEV_1/FVC percentage predicted values were positively correlated with log_2 methacholine PC_{20}. They also found that the $FEF_{25\%-75\%}$ percentage predicted values correlated better with log_2 methacholine PC_{20} and bronchodilator responsiveness than FEV_1/FVC percentage predicted values or FEV_1 percentage predicted values. From the performed ROC curve analysis, it was found that the $FEF_{25\%-75\%}$ at 65% of predicted value had a sensitivity of 90% and a specificity of 67% for the

detection of a increase of 20% in FEV_1 after inhalation of albuterol. They conclude that $FEF_{25\%-75\%}$ percentage predicted values should be evaluated in clinical studies of asthma in children, and that it might be useful in the prediction of the presence of clinically relevant reversible airflow obstruction.

Bacharier et al. (2004) [46] researched if lung function measures are consistent with asthma severity. Parents of 219 children in the age range of five to eighteen years old were asked to fill in a questionnaire about asthma medication and symptom frequency in the last one and four weeks. Next to that the children performed spirometry. It was found that FEV_1 percentage predicted value and FVC percentage predicted value did not differ by level of self-perceived asthma severity when this severity level is based upon symptom frequency or medication use, or a combination of those, while the ratio FEV_1/FVC and $FEF_{25\%-75\%}$ decreased significantly when asthma severity increased. From this they concluded that FEV_1/FVC decreases when asthma severity increases while FEV_1 stays normal. However, discriminant analysis showed that when these values were used to classify patients in a severity category, the FEV_1 percentage predicted value classified 33% correct, the FEV_1/FVC ratio 32%, and the $FEF_{25\%-75\%}$ 39%. There were four categories, making the prediction not much higher than predictions based on chance (25% per category). The variability of $FEF_{25\%-75\%}$ was twice as much as the variability of the FEV_1/FVC ratio.

The goal of research performed by Ratageri et al.(2001) [47] was to find out which lung function measurements resulted in a better assessment of asthma severity. Sixty children in the age of five to fifteen years old without asthma, and with mild or severe asthma, were studied. A portable spirometer was used to measure several values. It was found that using FEV_1 and FVC, children with mild asthma could be differentiated from children without asthma in 63% and 58% of the cases. $FEF_{25\%}$ was able to identify 77% of the cases and $FEF_{75\%}$ 90% of the cases. In the group with children diagnosed with severe asthma FEV_1, FVC, $FEF_{25\%}$, and $FEF_{75\%}$ were abnormal, compared to children without asthma, in respectively 90%, 80%, 97%, 94% of the cases. PEF was found to be abnormal in 77% of the mild cases and 87% of the severe cases. When looking at the FEV_1/FVC ratio no significant difference was found between asthmatic and non-asthmatic children. From these results it is concluded that the $FEF_{25\%}$ and $FEF_{75\%}$ are more useful measurements for the assessment of asthma severity than FVC and FEV_1. Additionally, the FEV_1/FVC is found to be not useful at all in this study.

Conclusion

Several conclusions can be drawn from the literature. It was found by several studies that PEF has a poor concordance with e.g. symptoms, FEV_1, and lung function [31, 33]. It was also found that PEF has no additional benefit over self-reported symptoms [36, 37, 39, 40]. People respond to symptoms instead of changes in PEF when adjusting their medication [38].

Also, none of the values symptoms, atopy, PEF, FEV_1, and PD_{20} should be used on their own as it does not give a complete overview of the lung function [32]. Besides, it was found that FEV_1 is a poor indicator of asthma severity [44], but MDV was found to be a useful indicator of asthma severity [34].

FEV_1 percentage predicted value was found to be an indicator of an asthma attack in the year following the measurement [43]. Nonetheless, it was also found that the FEV_1 percentage predicted value did not differ by level of asthma severity when this level is based upon symptom frequency or medication use, or a combination of those [46].

Another conclusion that can be drawn is that the $FEF_{25\%-75\%}$ is a more sensitive indicator of airflow obstruction compared to PEF, FEV_1, FVC, FEV_1/FVC ratio and symptoms. A given reason is that it is independent of effort. Besides, it measures the airflow in the peripheral airways and there is a high probability of persisting airway obstruction even when the patient has no symptoms and a normal PEF [35]. $FEF_{25\%-75\%}$ decreases significantly when asthma severity increases [46]. Simon et al. (2010) [45] state that $FEF_{25\%-75\%}$ percentage predicted value is a good indicator.

Ratageri et al. (2001) [47] stated that FEV_1/FVC ratio is not useful at all. This is in conflict with Ramsey et al. (2005) [41] who mention that FEV_1/FVC is a useful ratio and Bacharier et al. (2004) [46] who state that this ratio decreases significantly when asthma severity increases. A reason for these differences can be the different manners in which the asthma severity was determined. In the research of Ratageri et al. (2001) [47], the asthma severity was based on guidelines of the international pediatric asthma consensus group [48]. The asthma severity in the research of Bacharier et al. (2004) [46] was based on medication and symptoms. Ramsey et al. (2005) [41] determined the asthma severity using a questionnaire which was in concordance with the National Athma Education and Prevention Program (NAEPP) guidelines. These differences in determining the asthma severity levels make the comparison dubious.

3.4 Spirometry and games

In order to improve spirometry measurements, the present project offers little games to the patients, also called metaphors, based on the frequently made errors by the patients, to support them during the following spirometry measurements. Researches that evaluated how games influences spirometry measurements are presented in this section.

Vilozni et al. (2001) [49] compares a one target candle-blowing game with the multi-target game SpiroGame which divides the spirometry test in three phases; full inspiration before expiration, instant forced expiration, and long expiration to residual volume. 102 children in the age range of three to six years were asked to perform spirometry using the two games, in randomized order. 69.6% of the children were able to produced acceptable spirometry using SpiroGame, against 47.1% of the children using the candle-blowing game. The main reason for failure was the same in both systems; poor effort due to lack of coordination or comprehension. 79% of the children using SpiroGame versus 4% of the children using the candle-blowing game were able to reach acceptable FEV_1. A premature expiratory break was seen in six children when using the SpiroGame, and forty-one using the candle-blowing game. FVC results were similar, and PEF was higher when using the candle-blowing game. The teaching time was comparable between the two games. It can be concluded that dividing the game in multiple targets, to address the multiple phases of the spirometry manoeuvre, increases the performance.

Vilozni et al. (2005) [50] also analyzed what the role of SpiroGame is in supporting spirometry. The participants were in the age range of two to 6.5 years old. 78% of the children was able to perform three acceptable measurements using SpiroGame. The reasons why measurements were not acceptable were mainly a lack of comprehension of the $FEFV$ maneuver. Thirteen children refused to play the game. This was similar to results from a study based on verbal coaching [51]. The difference was that PEF values and flow-related volumes were higher when using the games. This could be due to the splitting of the manoeuvre in different targets, clarifying every step visually and audibly.

Research done by Gracchi et al. (2003) [52] focuses on the question if adding computer animated programs to the spirometry procedure improve the results of spirometry sessions. Eighty-eight children in the age group of four to eight years were involved in this research who all performed two series in which at least three acceptable curves were produced. One series was done with the computer animated program and one without, in a randomized order. The computer animated programs used two games; one in which the participant is asked to blow out five candles which is triggered by the peak flow and one in which they are asked to

blow up balloons, also triggered by the peak flow, after which they have to keep the balloons in the air as long as possible, which is triggered by the FVC value. They found that when using the computer animated programs, less participants were able to create an acceptable FVC and FEV_1, but a better PEF was seen. This could be due to the fact that the games focus on reaching a good PEF, and the children stop putting in effort when this target is met. In the age group six to eight years old, the performance of FVC decreased significantly.

Kozlowska et al. (2004) [53] present a critical note on this research. At first, they believe the programs were not used to their full potential. Secondly, they believe the reproducibility criteria are too strict. They agree with the finding that the use of these incentive programs does not offer much advantage for the age group of six to eight years old, however they believe that the younger children may gain advantage from it as it is more difficult for them to understand, process and carry out multiple steps. Another interesting point they note is that it may be the case that the target used in this research can be underestimated as it is probably based on too little data. Therefore, the children do not use their full potential during the measurements.

Gracchi et al. (2003) [52] reacted to this critical note and mentioned that they see the advantage for younger children, however they were not able to prove that it was helpful for improving reproducibility and maximal effort. They did some research to the reference values used as targets, which were based on data from older children, and concluded that this indeed may be a problem for the FEV_1 value, but not for the FVC. When looking at their results they state that this did not influence their results, however the use of higher targets should be studied.

Conclusion

Vilozni et al. (2001) [49] show promising results using a multi-target game; 69.6% of the children was able to produce an acceptable measurement, compared to 47.1% when using the candle-blowing game. Gracchi et al. (2003) [52] showed a negative result as the quality of FEV_1 and FVC declined. However, the experiment was not performed right, according to Kozlowska et al. (2004) [53]. A finding of Gracchi et al. (2003) [52], which is agreed upon by Kozlowska et al. (2004) [53], is that the incentive programs do not offer much advantage for the age group of six to eight years old, however younger children may gain advantage from it as it is harder for them to understand, process, and carry out multiple steps.

The age of the target group of the present project is six to eleven. As the children participating in the aforementioned experiments were not all in this age group, it makes it hard to apply these results directly. Instead, the present project will provide additional insights.

3.5 Inter- and intra-annotator agreement when assessing (errors in) spirometry data

To interpret the results from the error detection algorithm, it is necessary to know what the agreement is in error detection in spirometry data when assessed by multiple professionals.

In research performed by Velickovski et al. (2018) [54], three clinical experts assessed 600 spirometry curves of adults, based on the flow-volume curve and the volume-time curve, to determine if it has to be rejected or not. A mean kappa score of 0.34 for the inter-rater agreement was found. The intra-rater agreement was not calculated.

Tuomisto et al. (2008) [55] determined the intra- and inter-rater agreement between two clinical psychologists. The first had thirty years of experience, while the second one had 20 years of experience. Both assessed curves of adults to evaluate if the start was without delay, if the steep was upslope, if the PEF was sharp, if there was no cough in the attempt, and if the exhalation was full. The intra-rater agreement was calculated over twenty-five curves per clinical psychologists and was 98% and 99%, respectively. Both assessed fifty curves over which the inter-rater agreement was calculated, which was found to be in the range from 83% to 100% when assessing the agreement for every criteria separately.

Inter-rater agreement is determined by Seyedmehdi et al. (2013) [56], based on 100 curves of adults which were assessed using a checklist of errors derived from the ATS/ERS criteria. This is done by two occupational medicine specialists. The found kappa coefficient of the inter-rater agreement was 95%. The intra-rater agreement is not calculated.

Conclusion

The inter-rater agreement ranges from 0.34 to 100%. The lowest inter-rater agreement was found when evaluating if a spirometry attempt has to be rejected based on the flow-volume and the volume-time curve. When assessing if specific criteria are met, or if errors are present in the spirometry attempt, the agreement was much higher, ranging from 83% to 100% [55, 56].

3.6 Conclusion literature review

The literature review covered several topics such as spirometry in children, the quality and compliance of spirometry at home, and the involvement of games in spirom-

etry, based on the questions stated in section 3.1.

The questions asked about spirometry in children were what the differences are, what these imply and how the consequences are dealt with. From the literature it could be concluded that it is necessary to use specific criteria for children when they are performing asthma manoeuvres as the lungs of children are less developed than lungs of adults. Examples are changing the criteria for an acceptable FEV_1 to 80 ml, or 12.5% of the FVC. Also, the expected FET should be lowered to less than three seconds. Another example is that the difference between the FVC and the FEV_1 needs to be within 100 ml, or 10% of the highest value, when looking at between-manoeuvre criteria.

The questions asked about spirometry at home were what the quality of spirometric data derived during home spirometry is, what procedure related factors influence this quality, and which values derived from spirometry data are useful in monitoring or diagnosing asthma. From literature it was concluded that the quality and compliance is satisfactory, however the compliance decreases after two to four weeks, and keeping a diary is still necessary next to spirometry tests, to create a complete view of the asthma condition. This diary should include symptoms and factors (e.g. facial pain), that can have an influence on the measurements. Using this diary to write down PEF results is found to be not reliable as entries are faked. Therefore, the combination of saving PEF values automatically, and using a diary to write down factors that probably have an influence on the measurement, is ideal. Encouragement during a measurement, and showing correcting instructions after a non-acceptable attempt, increased the quality assurance of the next attempts. The main errors made during the spirometry test were abrupt ending, invalid time to PEF, variable effort in combination with cough and glottis closure. The most useful value to take into account when monitoring or diagnosing asthma is $FEF_{25\%-75\%}$; PEF and FEV_1 are found to be not useful, the usefulness of the ratio FEV_1/FVC is unclear.

The question asked about spirometry and games is how games used during spirometry attempts influences the spirometry quality. When looking at the effect of games on spirometry measurements, good results were found; Vilozni et al. (2001) [49] found that 69.6% of the children using the game SpiroGame were able to produce an acceptable measurement compared to 47.1% using the candle-blowing game. Only Gracchi et al. (2003) [52] showed a negative result; the quality of FEV_1 and FVC declined in his experiment. However, according to Kozlowska et al. (2004) [53], the experiment was not performed right as amongst other things the PEF target was probably too low.

The inter- and intra-rater agreement when assessing (errors in) spirometry data was explored. It was found that the inter-rater agreement for assessing the errors

in spirometry data of adults is high, ranging from 83% to 100% [55, 56]. However, Velickovski et al. (2018 [54] found a poor inter-rater agreement of 0.34 when the clinical experts were asked to state if a spirometry attempt should be rejected or not. The intra-agreement when assessing errors in spirometry data found by Tuomisto et al.(2008) [55] was high; 95%. These researches show that the difference in assessing errors in spirometry data is small between professionals, and for one professional over time, however high when assessing if an attempt has to be rejected or not.

To summarize, unsupervised spirometry at home is possible for a period of two to four weeks, after which the compliance decreases, in combination with keeping a diary to keep a complete overview of the condition of the patient. If the patients performing spirometry are children, adapted criteria should be used and the most useful value to take into account when monitoring asthma is $FEF_{25\%-75\%}$ while PEF and FEV_1 are found to be not useful. The addition of games to the spirometry measurements result in higher quality spirometry data. Besides, the agreement when assessing (the errors in) adult spirometry data between professionals, and for one professional over time, is most of the time very high.

Chapter 4

Related work

One of the goals of the present study is designing an error detection approach using machine learning to detect errors in spirometry data. This chapter discusses previous work in this area.

4.1 Method related work

A systematic approach is used to select relevant papers. Several search terms were used:

1. spirometry error detection
2. spirometry error detection machine learning
3. spirometry machine learning
4. spirometry errors machine learning

The first search term did not give any useful results, using the second term resulted in four papers which seemed useful as the title and/or abstract mentioned ”machine learning” and ”spirometry”. Not all papers mentioned ”errors” or something alike. The reasoning was followed that although it was not directly about errors, if it was about machine learning and spirometry, it could still be useful to read. When search term three was used, two other papers were found which stated ”spirometry” and ”machine learning” in their title or abstract. Using the last term, three more papers were found which stated ”spirometry” and ”machine learning” or a machine learning technique in their title or abstract. Besides using the search terms, the references of the selected papers were assessed to select papers describing related work as well.

Unfortunately, not all papers which seemed promising when selected were useful; a lot of papers did not use spirometry data in the end as input for the machine

learning models, but e.g. results from CT scans. Besides, some papers were about predicting FEV_1 values instead of using these values to predict the asthma severity. The papers which used spirometry data as input and were about error detection in spirometry data or about the diagnosis of asthma from spirometry data are described in this section. Although this project focuses on the error detection and not the diagnosis, knowing related work in this area is still useful as apparently the combination of input and technique reflected the asthma severity of people.

Using this method, four useful papers were selected. Only one paper was found about detecting errors in spirometry data using machine learning. This shows that the present study is a potentially big contribution to known research.

The following sections discuss the selected papers.

4.2 Error detection

As said, not much research has been performed in the automatic detecting of errors in a spirometry attempt. Luo et al. (2017) [57] did an attempt to detect four common errors by creating a separate classifier for every error. The data consisted of curves from patients in the age range of three to ninety-five of which 72.2% came from patients between six to eight-teen years old. Curves derived from spirometry data were manually labeled by six professionals. The curves were all labeled by one professional and were not compared in an inter-annotator agreement analysis. The classifiers were trained for each error using between 1314 and 5728 curves with this error as positive cases, and an equally-sized random sampling dataset without errors as negative cases. 90% of the data was used for training, the rest for testing. sixty-eight features were selected based on feedback by doctors, spirometry training materials, and based on related work. However, it is unclear which features are used. The fact that not all curves are labeled by one person leads to a potential for noisy labels. For this reason an ensemble method was used. An Ada-Boost classifier with decision trees was used as the base estimator. The precision, recall, and F-score were the evaluation metrics used. All the features derived from the data were used as input, resulting in an F-score of 0.92 for early termination, 0.86 for detecting a cough, 0.86 for detecting variable flow, and 0.85 for detecting extra breath. The most significant features for detecting early termination were the total time which elapsed during an attempt, and the volume which was exhaled in the last second of the attempt. For detecting a cough, the most significant features were the maximum slope found in the $FV\ curve$ after the peak flow, and a heuristic for the total amount of time the slope in the $VT\ curve$ is relatively flat. This heuristic is found by looking at the period of volume exhaled where the slope of the $FV\ curve$ is less than 10% of the maximum slope. The significant features for detecting an extra

breath taken were the minimum slope in the $VT\ curve$, and the maximum slope in the $FV\ curve$ after the peak flow. For detecting variable flow, the most significant features were the maximum slope in the $FV\ curve$ after a peak flow, and the sum of the total first derivative whose values were positive after the area of the highest flow in the $FV\ curve$.

From this we can conclude that the detection of early termination, cough, variable flow, and if an extra breath is taken is possible using an AdaBoost classifier with decision trees using the mentioned features. The authors believe that using RNNs, as they are able to recognise pattern in time series data, would be an interesting approach for future research as this would remove the need to construct features manually.

4.3 Diagnosis of asthma

Research has also been done in the field of diagnosis of restrictive spirometric patterns or airway obstruction by using machine learning techniques. Sahin et al. (2010) [58] uses multi-class SVMs to predict the diagnosis of spirometric patterns in the classes normal, restrictive, or obstructive. The decisions of the SVM were fused by using error correcting output codes ($ECOC$). This multi-class SVM combined with ECOC was trained on FEV_1, FVC, and the FEV_1/FVC ratio. 499 measurements produced by male subjects between twenty-five and forty-five years old were used as input. These measurements were according to the guidelines of ERS. The trainingset included 162 normal measurements, twelve restrictive, and twenty-six obstructive. The testset included 246 normal measurements, eight-teen restrictive, and thirty-five obstructive. To choose the right kernel function for the SVM, several functions were empirically studied and it was found that the radial basis function (RBF) with a σ value of 0.3 gave the best results. The C value was also found by trying out different values and it was found that a value of eighty gave the best results. However, for both values it is unclear what the empirical study entails exactly. Several optimization techniques were used during training, such as decomposition and caching. It is unclear which other optimization techniques are used. The specificity, sensitivity, ROC curve, and accuracy were used as evaluation metrics. The found total classification accuracy was 97.32%. The specificity was 97.97%, the sensitivity for the restrictive class was 94.44%, and 94.29% for the obstructive class. The normal patterns and the obstructive patterns were most often confused with restrictive patterns. These results show the usefulness of using SVMs in the diagnoses of spirometric patterns.

In research performed by Bright et al. (1998) [59], neural networks are used to detect upper airway obstruction (UAO). Data from 155 adults of which forty-six

probably have UAO, fifty-one without UAO, and fifty with airflow limitation, which is caused by $COPD$, was used as input. The used data consisted of one curve per person and was acceptable according to the recommendations of the British Thoracic Society and the Association of Respiratory Technicians Physiologists [16]. The curves were examined twice by two professionals. The intraobserver kappa score was 0.86 for observer one, and 0.63 for observer two. The interoberserver kappa score was between 0.58 and 0.68 for each of the classification sessions. The features used were PEF, FEV_1, FVC, FEV_1/FVC ratio, and the FEV_1/PEF ratio. Besides, the flatness of the curve is taken into account as the FV-curve is relatively flat in the early part of the curve in subjects with UAO. Another feature is the moment ratio[1]. The last feature is the FEF_{50}/FIF_{50} ratio were FIF is the Forced Inspiratory Flow. This feature is not calculated for the curves of the patients with COPD as this data was not available. The network had two hidden layers; it was found that approximately twice the number of input nodes for the first layer, and half the number of input nodes for the second layer produced the best classification. How this is determined is unclear. The dataset which was used to train the models was randomly selected by the program which was one half to two thirds of all curves from every group of subjects. The test set was a separate set of data. The model is trained multiple times using different training sets. The evaluation metric used is Cohen's kappa statistic. An α of 0.5 was used to determine the significance. Four different networks are trained. The first network has as input all features except for the FEF_{50}/FIF_{50} ratio, the second network included this ratio, the third network used the five inputs with the highest relative contributing factors from the first neural network together with the FEV_1/PEF ratio, and the fourth network was the same as the third but without the FEV_1/PEF ratio. To calculate the contribution factor, the weights of all neurons from a particular input to its output were summed. Next to the neural networks, two logistic linear regression models were developed; one using the same inputs as the third neural network, and one using the same input as the fourth one. To avoid overfitting, the training was terminated when the error rate for detecting UAO was at a minimum in both the training set and test set. It was found that a combination of the flatness of the expiratory loop, the FEV_1/PEF ratio, and the moment ratio obtained the best results; this resulted in a sensitivity of 88%, a specificity of 94%, and an accuracy of 92%. The flatness score resulting in the best classification were the length of the chords taken at 95% and 75% of the PEF, and at 1.5 and 2.0 L/s below PEF. This information was not sufficiently effective on its own, however improved the classification compared to only using the FEV_1/PEF

[1]The moment ratio is calculated using the mean of the transit times per milliliter (α_1), and the mean of the square of all transit times after standardizing these by the expired volume (α_2). The moment ratio used as a feature is calculated as follows: $MR = (\alpha_2)^{\frac{1}{2}}/\alpha_1$.

ratio as input.

Another research which uses neural networks to classify spirometer data is performed by Manoharan et al. (2008) [60]. This research focuses on the use of two different Artificial Neural Networks (ANNs); one with a radial basis function (RBF), and a back propagation neural network. Data from 150 participants, hundred for training, and fifty for testing, was used in this research; twenty-five obstructive, twenty-five restrictive, fifty normal, and fifty for validation. This data consisted of one curve per person and was acceptable according to the ATS criteria. The features obtained from the data were FVC, FEV_1, FEV_1/FVC%, PEF, and $FEF_{75\%}\%$. Also, the predicted and percentage predicted values of these features were feeded to the neural networks. The rest of the preprocessing process is not described. The feed forward neural network used has one hidden layer using a log sigmoid transfer function. It is unclear how the decision is made for the amount of hidden layers and transfer function. The radial basis function neural network is a multi-layer feed forward network which consists of one hidden layer of non linear units which operate as kernel nodes. The output layer has linear weights. The activation function for the hidden nodes was a radially symmetric Gaussian radial basis function. The performance was evaluated using the accuracy, sensitivity, specificity, false positive rate, positive predictive value, negative predictive value, and adjusted accuracy [2]. It was found that the RBF neural network is more sensitive in comparison with the back propagation neural network with an accuracy of 100% versus 96%, an equal sensitivity of 100% , a specificity of 100% versus 92.59%, a false positive rate of 0% versus 7.41%, a positive predictive value of 100% versus 92%, an equal negative predictive value of 100%, and an adjusted accuracy of 100% versus 96.30%. However, it is mentioned that with a larger database and more features the back propagation network could be enhanced.

4.4 Discussion

Several options for detecting errors in spirometry data or diagnosing lung diseases, using spirometry data, are discussed. However, the descriptions of the performed research are not complete in all addressed papers. Parts of the description of the preprocessing process are missing. Examples are the balancing method used, and the calculations of the features. The features itself are described in all papers, except

[2] Sensitivity = $TP/(TP+FN)$, Specificity = $TN/(TN+FP)$, False positive rate = $FP/(TN+FP)$, Positive predictive value = $TP/(TP+FP)$, Negative predictive value = $TN/(TN+FN)$, Adjusted accuracy = $sensitivity + specificity)/2$
Where: TP = True Positive values, TN = true Negative values, FP = False Positive values, FN = False Negative values

in the paper by Luo et al. (2017) [57]. Nonetheless, all papers do not motivate the choices for the features chosen. This could for example be very helpful in the paper by Sahin et al. (2010) [58] where only three features are used.

Next to the preprocessing phase, the model optimization phase is not explained in detail in the papers. Bright et al. (1998) [59], Sahin et al. (2010) [58], and Manoharan et al. (2008) [60] do not describe why a certain model is chosen. Additionally, Sahin et al. (2010) [58] mentions two optimization techniques, but not the complete list of techniques used. Knowing these techniques would be very helpful in designing our SVM. The paper describing the research performed by Bright et al. (1998) [59] does not describe which activation functions are used and why. The paper by Manoharan et al. (2008) [60] does not describe which activation function is used and how many hidden layers are used together with the reasoning.

The performance metrics used are explained very well in the papers. However, the paper by Manoharan et al. (2008) [60] misses an evaluation of the results. The results are very high, with a sensitivity in both cases of 100%. This is almost impossible using machine learning and so a critical evaluation of the results would have been in place.

4.5 Conclusion

In conclusion, several options have given good results in previous work. An AdaBoost classifier with decision trees can be used to detect 4 errors: early termination (F-score: 0.92), using the total time of an attempt and the volume exhaled in the last second of the attempt, cough (F-score: 0.86), using the maximum slope after peak flow and a heuristic for the total amount of time the slope in the $VT\ curve$ is flat, variable flow (F-score: 0.86), using the maximum slope in the $FV\ curve$ after peak flow and the sum of the total first derivative whose values were positive after the area of the highest flow in the $FV\ curve$, and if an extra breath is taken during the attempt (F-score: 0.85), using the minimum slope in the $VT\ curve$ and the maximum slope in the $FV\ curve$ after peak flow. Additionally, the suggestion is made to use RNNs to detect errors in a spirometry attempt.

Options given to diagnose asthma are multi-class SVMs trained using the FEV_1, FVC, FEV_1/FVC ratio as features and a radial basis function as kernel function, a Neural Network using the flatness of the expiratory loop, FEV_1/PEF ratio, and the moment ratio as features, and an ANN using a radial basis function and FVC, FEV_1, FEV_1 %, PEF, $FEF_{75\%}$ as features. Although this is about the diagnosis of asthma, it shows that correlations are found between the aforementioned values and asthma severity when these techniques are used.

In chapter 6 the approach of the present study based on these results is shown.

Unfortunately, the researches were not described in detail, and only one paper about error detection in spirometry data was found, making it hard to apply these approaches directly. Therefore, the method is inspired by these approaches but complemented with own knowledge and insights from professionals.

Chapter 5

Research Questions

As explained in chapter 1, this research focuses on three parts of the project SpiroPlay; designing and evaluating an error detection approach using machine learning, evaluating the agreement in error detection between professionals, and the evaluation of using metaphors as a coaching manner on the quality of spirometry data.

The first research question is about the to be designed error detection approach. At the moment, professionals visually observe the person performing the measurement and the resulting flow-volume curve, to determine if and which error(s) occurred. However, this means that a professional needs to accompany the patient while the measurement is conducted which should not be the case during home monitoring. The error detection approach aims to support the professional by identifying the errors without intervention of a professional.

The machine learning approach will be compared to the rule-based error detection approach used in the SpiroPlay system nowadays, to evaluate which approach is advised to use in the SpiroPlay system.

The following research question and subquestions are answered in this research:

RQ 1: How well can an error detection approach using machine learning techniques detect errors in spirometry data?

SQ 1.A: Which procedure based on the most promising procedures found in literature is able to detect errors most accurate?

SQ 1.B: How does the machine learning approach perform compared to the rule-based approach designed by V. De With?

To be able to interpret the answers to the first research question, the consistency in detecting errors in spirometry data by multiple professionals need to be determined. If professionals are not consistent in this, it shows that the detection

of errors in spriometry data is ambiguous. An inter-annotation study, explained in chapter 6, is set up to answer the following research question and subquestions:

RQ 2: What is the agreement in detecting errors in spirometry data by professionals?

SQ 2.A: What is the agreement between two professionals in detecting errors in spirometry data?

SQ 2.B: What is the agreement in detecting errors in spirometry data by one professional over time?

The third research question focuses on the influence of the metaphors on the quality of the spirometry measurements. If the quality is comparable or increases when using the metaphors compared to when the children are coached by a professional, it shows that the metaphors are a decent coaching manner to coach children during home spirometry.

The following research question and subquestions are answered in this research:

RQ 3: What is the difference in quality of the spirometry measurements between blowing behaviour coached by a professional and coached by a metaphor offered by the SpiroPlay system?

SQ 3.A: What is the difference in measured PEF between blowing behaviour coached by a professional and coached by a metaphor offered by the SpiroPlay system?

SQ 3.B: What is the difference in measured FEV_1 between blowing behaviour coached by a professional and coached by a metaphor offered by the SpiroPlay system?

SQ 3.C: What is the difference in measured FVC between blowing behaviour coached by a professional and coached by a metaphor offered by the SpiroPlay system?

SQ 3.D: What is the difference in the number of errors between blowing behaviour coached by a professional and coached by a metaphor offered by the SpiroPlay system?

The approach to answer these research questions is elaborated in chapter 6. The answers to the research questions are presented in chapter 9.

Chapter 6

Method

This section elaborates the approaches used in this research to answer the questions stated in chapter 5. The first section describes the implementation and evaluation of the error detection method, the second section describes the inter-annotation study, and the third section the method of analysis of the effect of the metaphors on the quality of spirometry tests.

6.1 Error detection

The first part of this research focuses on detecting the errors in spirometry data. The process from data gathering to evaluation consists of different steps, which are visualised in figure 6.1. The blue boxes indicate the data preparation steps, namely the data gathering, preprocessing, and data segregation, whereas the orange box indicates the training step, and the green box indicates the evaluation step.

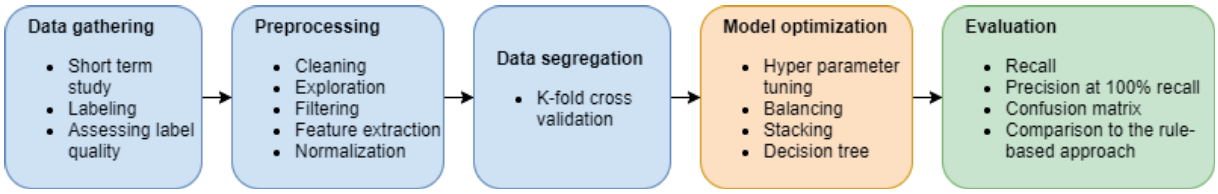


Figure 6.1: The machine learning pipeline of this study

These steps are elaborated further in the remainder of this section.

6.1.1 Data gathering

Data gathering is performed using the spirometer (refer to section 2.3) and an app which utilizes the rule-based evaluation approach, designed by V. De With (refer to section 6.1.4 for details). Thirty children were asked to perform two test; one

with a metaphor, and one without. One test consists of three or more attempts with a maximum of eight. During the test, the attempts are labeled by a professional indicating which errors occurred, using the form presented in Appendix D.1.This professional has experience with assessing spirometry curves from asthma patients. From her expertise she knows how errors look like by visually observing the curves and the patient during the test. The list of errors ordered in order of importance, are shown in table 6.1 together with their code. These errors are a subset of the criteria set by Miller et al. (2005) [3], which are described in section 2.2.1.

Additionally, the data collected during the label quality experiment, explained in section 6.2, is used to assess the effect of including more data in the dataset. However, as this is data from spirometry attempts performed by adults instead of children, it is not on forehand obvious including this data will increase the performance.

The total amount of attempts collected during the hospital experiments is 309. The data from the inter-annotation study consists of 191 attempts. Combining both datasets will sum up to a total of 500 attempts. This is not much data, but as the data is collected by experiments involving human beings, it is difficult to collect a large amount of data.

Error	**Code**
Extrapolated volume $<$5% of FVC or less than 0.15 L at the start of the expiration	1
Obstructed mouthpiece	2
An extra breath taken during the attempt	3
Flow leak	4
Duration of $<$3 s, a plateau in the VT curve, or if the person can/should continue exhaling	5
No maximal effort	6
Cough	7
Wrong posture	8
Other, namely:	9
No error	0
Uncertain	-1

Table 6.1: Complete table of error codes

6.1.2 Preprocessing

Before the data can be used in model training, it has to be preprocessed. This includes cleaning and exploring the data, feature extraction, and normalization.

Cleaning the data

The first step in the preprocessing phase is to clean the data. Samples with label ”uncertain” were deleted as these serve no purpose in training, testing, and validation.

Furthermore, the inhalation and exhalation time window was determined for each attempt. The parts of the data before inhalation and after exhalation is noise and therefore removed. The start of the inhalation is where the attempt is zero for the last time before the exhalation curve. If this is never the case, the start of the inhalation is the same as the start of the data. The end of the exhalation is where the exhalation curve is above 0.1 for the last time. This threshold is determined by the professional who also labeled the data, by empirical tests with the used spirometry device.

Figure 6.2 shows an example curve, with the inhalation and exhalation parts annotated.

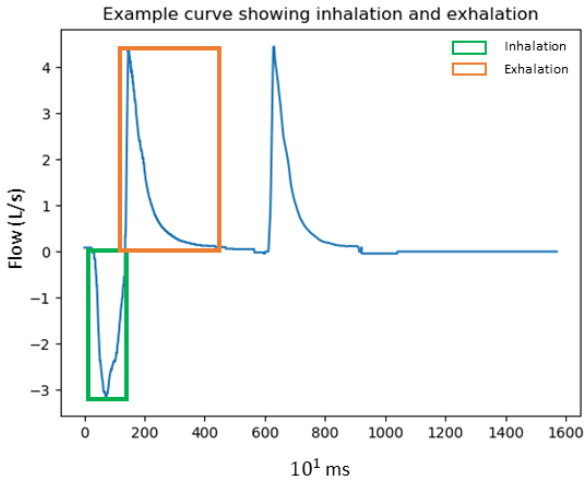


Figure 6.2: An example flow-time curve.The inhalation and exhalation parts are annotated. The rest of the data is noise.

Additionally, the curves were smoothed to reduce noise, using an Savitzky-Golay filter (refer to appendix A.3 for an explanation).

Data exploration

In the data exploration phase, the data is visualised. First, a histogram of the detected errors is created, in order to evaluate if the data is imbalanced, and if all errors presented in table 6.1 are covered in the data.

Secondly, a correlation plot is created to appraise the extend to which the different errorclasses are linearly related.

Filtering

Outliers are removed to filter the data and to reduce noise. Outliers were found by visually assessing the attempts. An attempt was seen as an outlier if it had bumps in the inhalation, which passes the zero line. An example can be found in figure 6.3.

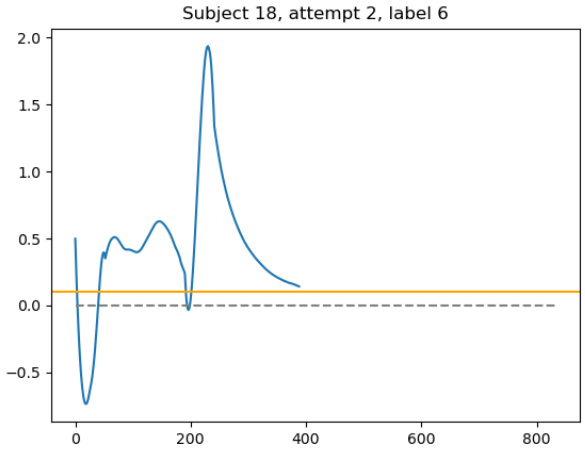


Figure 6.3: An example of an attempt seen as an outlier by visual inspection.

Next to assessing the attempts visually, the means of the FVC and FEV_1 values were compared on a per person basis. If a value was $\geq$ 2.5 standard deviations away from the mean value, the measurement was removed from the dataset.

Feature extraction

The final step of the preprocessing phase is to extract features which are used as input for the machine learning models. Four different featuresets were extracted from the data. The featuresets and their combinations are clarified below.

Spirometry parameters
A lot of parameters can be extracted from spirometry data, such as the FVC, FET, and FEV_1 (refer to section 2.2 for details). These parameters are found to be useful in diagnosing asthma [58–60]. However, it is unclear if these features will be useful in detecting errors in spirometry attempts as well. Therefore, several spirometry parameters are extracted from the data and used as input for the to be trained models. The spirometry parameters extracted are presented in table 6.2.

Feature	Explanation
FVC	Total volume
$FEV_{0.5}$	Volume in the first half second
FEV_1	Volume in the first second
PEF	The flow on the highest peak
EV	The extrapolated volume
$FIVC$	The Forced Inspirational Vital Capacity
FEV_1/FVC *ratio*	Ratio of volume in the first second to the total volume
FET	Total time
FEF_{25-75}	Volume between 25% and 75% of FVC

Table 6.2: The spirometry parameter featureset.

Time-series features, unfiltered

Detecting errors in spirometry data is expected to be based on time-series features, for example cough is characterized by having two peaks in the spirometry data instead of one. This information is not represented by the spirometry parameters. Therefore, seventy-two time series features, such as the location of the maximum, and the number of peaks, were extracted from the time-series data. The complete list of time-series features can be found in appendix B. These features are extracted by using the tsfresh package for Python [1] as this method is recommended by Luo et al. (2017) [57].

Time-series features, filtered

The downside of extracting seventy-two features is that a lot of irrelevant features will probably be extracted as well which may influence the model negatively. To overcome this, the list of features will be filtered using the Feature Extraction based on Scalable Hypothesis tests (fresh) algorithm [61]. This algorithm filters the list of features with respect to their significance, but also takes the expected percentage by the machine learning models of selected but irrelevant features into account. The expectation is that this will create a featureset with features that explains the difference between the different error classes best.

Age and sex of the subjects

The age and sex of the subjects are of importance, as a child of, for example, eleven years old has a much larger lung volume than a child of six years old.

[1] https://tsfresh.readthedocs.io/en/latest/text/introduction.html

	Unfiltered timeseries features	Filtered timeseries features	Spirometry parameters	Age and sex
Unfiltered time-series features	X			X
Unfiltered time-series features + spirometry parameters	X		X	X
Filtered time-series features		X		X
Filtered time-series features + spirometry parameters		X	X	X
Spirometry parameters			X	X

Table 6.3: The featuresets used as input for the machine learning models.

Next to using the featuresets seperately, the featuresets will be combined to use as input for the models to be trained. All (combinations of) featuresets are shown in table 6.3. As the age and sex are important to distinguish between, for example, a low FEV_1 as an error and just the child being young and thus having a small lung volume, the age and sex are always used as features.

Normalization

The next step is to normalize the featuresets as most machine learning models assume all features are centered around zero and have a variance in the same order. If this is not the case, the features with a large variance might dominate the training process incorrectly.

Normalization can be done in various different ways. The two most used techniques in machine learning are min-max normalization and z-normalization. In min-max normalization the data is scaled so that every feature value is between zero and one. When using z-normalization, the data is transformed to have a mean of zero and a standard deviation of one. Refer to appendix A.2 for the formulas.

Both approaches have advantages and disadvantages. Min-max normalization is more robust to small standard deviations of features, and is not influenced by the scale of the features. However, the scale of the features is deleted which could be of importance.

Z-normalization is useful when working with different units or scales, and its output is closer to the expected input of the machine learning models than the output of min-max normalization, as the mean is zero and the variance is one.

Since the features used have different units, z-normalization will be used.

Generation of labelsets

As the dataset is sparse, it can be the case that some errors are not represented enough in the data to train a model on. Therefore, errorclasses are grouped to increase the amount of datapoints per class. It is expected that this will result in a better performance. The errorclasses are grouped in two ways. The errorclasses are divided in two classes; all attempts labeled with an error are now labeled as '1', all the attempt labeled as technically correct are now labeled as '0'. Although the model used for this classification is not able to detect which error is in the attempt, it is already valuable to know if there is one. The second way of grouping is combining all errorclasses which were labeled as 'other, namely: ', as these are not stated as a criteria by Miller et al. (2005) [3], and are therefore less important to be detected as a single error. The classifier used for this classification is not able to detect all single errors, but if it detects the attempt belongs to the class with the combined errors, general feedback can be given.

The grouping of errorclasses result in three different labelsets, which are presented in table 7.2.

The performance of the models when adopting the different labelsets is evaluated to appraise what is achievable in error detection in spirometry attempts.

6.1.3 Data segregation

'Leave-one-subject-out' K-fold cross validation is used to test the performance of the machine learning models on an independent dataset, to evaluate how well the model predicts unseen data. For background information on the standard K-fold cross validation approach one can refer to appendix A.4.

In the version of the K-fold cross validation approach used in this research, the test set always exists of all the samples from one subject. This implies K being equal to the number of subjects. Every participants performs two tests, which sums up to a minimum of six and a maximum of sixteen attempts per subject, and thus per testset.

Leave-one-subject-out cross validation is used to mimic the real life situation best by not including data from a test subject in the training data, as in real life, the attempt from a new patient is evaluated based on data from other patients without including data from the new patient.

6.1.4 Evaluation

The training and test sets were created by leave-one-subject-out cross validation,as explained in section 6.1.3. During hyperparameter tuning, Stratified K-fold cross-

validation is used, as explained in section 6.1.5.

The models were evaluated by the precision, recall, F1-score, and the precision at 100% recall. Refer to appendix A.9 for an explanation of these metrics.

Precision shows the classifiers exactness, recall the classifiers completeness, and the F1-score is the weighted average of the two. The metrics are not influenced by an imbalanced dataset while, for example, accuracy is.

The primary performance metric used is the recall, as it is most important that the model classifies the data as complete as possible. It is better to label attempts too many times as a certain error, than too little, as no useful feedback can be provided, and the attempt will not be useful in monitoring the asthma status of the patient.

Additionally, the precision at 100% recall, obtained from the precision-recall curve, is used as the secondary performance metric, as this metric represents the trade-off between precision and recall, using the recall as leading performance metric.

The performance metrics are designed for binary classification. To use these metrics in multiclass settings, the metric is calculated per class, and a weighted average is taken to prevent little classes from having a large impact on the final score. As in binary classification, the zero errorclass, representing attempts without an error, is neglected in averaging, as the only consequence of misclassification is to repeat the attempt. This does not result in quality loss, while it would be when misclassifying an attempt with an error as not containing one.

Next to the named performance metrics, the confusion matrices (see appendix A.9 for an explanation) are used to determine which error classes are confused when classified.

Comparison to a rule-based approach

The best model found in model training will be compared to a rule-based approach designed by V. De With. This error detection approach is based on evaluating the curves in a rule-based fashion. An example of a rule is that the first peak should be the only peak. If this is not the case, there has been, for example, a cough during the attempt. The implemented rules are based on a subset of the criteria given by Miller et al. (2005) [3], i.e. criteria one, three, four, five, six, and seven of the list with criteria stated in section 2.2.1. The rule-based approach is not designed to detect all errors, but focuses on the errors 'Unsatisfactory start' (error code 1), 'An extra breath taken during the attempt' (error code 3), 'Terminated too early' (error code 5), 'No maximal effort' (error code 6), 'Cough' (error code 7), and 'Glottis closure'. When a rule is not met, several types of errors can have occurred. An overview of the rules and which error codes are flagged when a rule is not met is presented in table 6.4. The presented error codes are the same as the codes in table 6.1, except

for glottis closure, which was not in the form used to file the errors.

Rule	Explanation	Error code
1	Examines if the extrapolated volume is smaller than 5% of the total volume, or 150 ml	1, 6
2	Examines if the first peak in the flow-time curve is the highest peak	3, 6, 7
3	Examines if the volume chances with less than 0.25 ml	5
4	Examines of the FET is larger than three seconds	5, glottis closure
5	Examines if there is no downward concave at the end of the flow-volume curve	5, glottis closure
6	Examines if the exhalation flow is constantly increasing until PEF is reached	7

Table 6.4: This table presents the rules of the rule-based approach, together with an explanation, and the error codes which occur when the rule is not met.

The rule based approach is used by the system during the hospital experiments to provide feedback to the children during the test with the metaphors, and to assess if three acceptable, and two reproducible curves were produced. When the child is coached by the professional, the rule-based approach also assessed the curves in the background, but the findings were not shared with the child during the test. The output from the system during the hospital experiments stating which rules were met, was used in the comparison.

The labels assigned to the attempts by the professional are used to compare both algorithms. The outliers and attempts with label 'uncertain' are excluded from the comparison. The output from the rule-based approach could not be directly compared to the labels of the professional, as not meeting a rule could imply different errors. Therefore, if the label given by the professional is one of the labels the rule-based approach could have assigned to the attempt, this label is used in the comparison.

When the label given by the professional did not overlap with the errors given by the rule-based approach, one error has to be assigned to the attempt to be able to compare the labeling by the rule-based approach and the professional. In these cases, the most important error, which is the error highest in table 6.1, is assigned to the attempt, as this would be the error chosen in real life to provide feedback for.

The precision at 100% recall could not be calculated for the rule-based approach, as only the labels are known and not the probabilities for every error per attempt. Therefore, this metric is excluded from the comparison, and the recall, precision and F1-score are used to compare the performance to the performance of the best

machine learning approach.

6.1.5 Model training

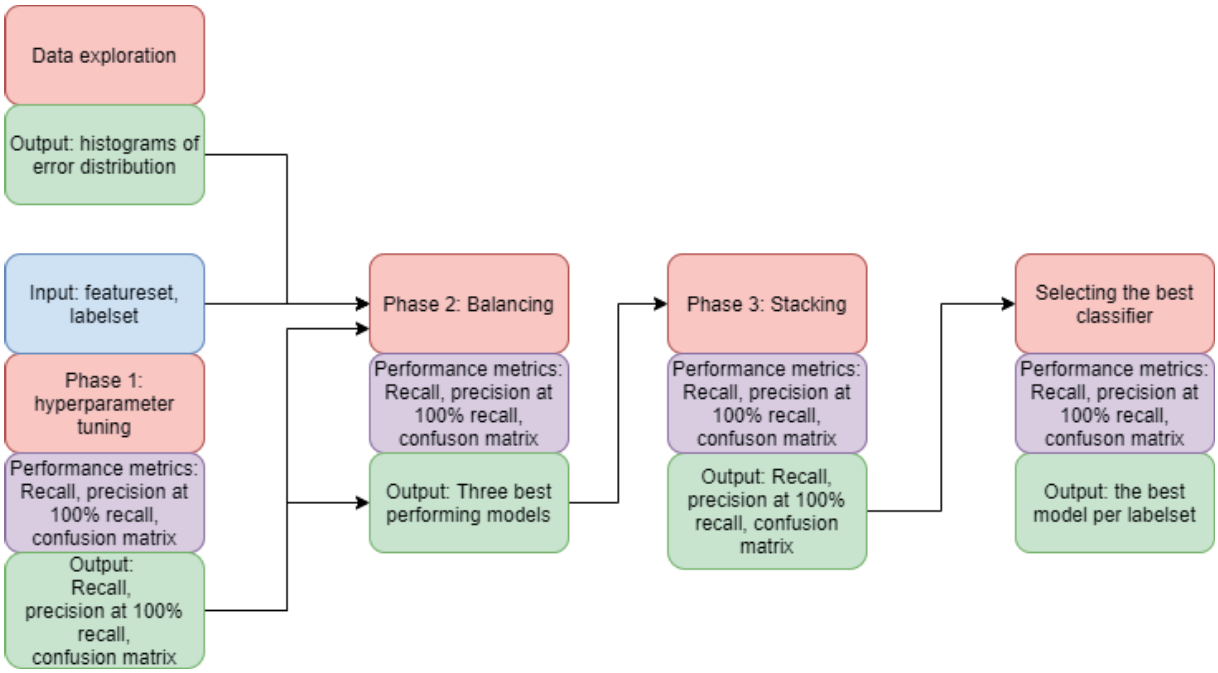


Figure 6.4: The phases of the model training and evaluation process.

Three phases were executed during model training, after which the best performing models were selected. These phases are presented in figure 6.4. This process was executed for every labelset. First, four different models with different hyperparameters, the different featuresets, and the different labelsets were trained. The performance of the models in this first phase, in combination with the histogram of the error distribution, were used to define the appropriate balancing technique for the second phase. After phase one and two, the best models per labelset are defined by the performance metrics explained in section 6.1.4. These best models are used in phase three where the models are combined to create a stronger model.

Next to finding the model which is best for predicting errors in the spirometry attempts for the three labelsets, a decision tree is proposed. This proposed decision tree is explained in detail at the end of this section.

Phase 1: Hyperparameter tuning

Different machine learning models were optimized and evaluated using programming language Python (version 3.7). A Long Short Term Memory model is trained

on the smoothed data while a Support Vector Machine, Radial Basis Function Neural Network, and boosted decision trees are trained on the different featuresets. All models were trained with the different labelsets. These models are explained in detail in appendix A.

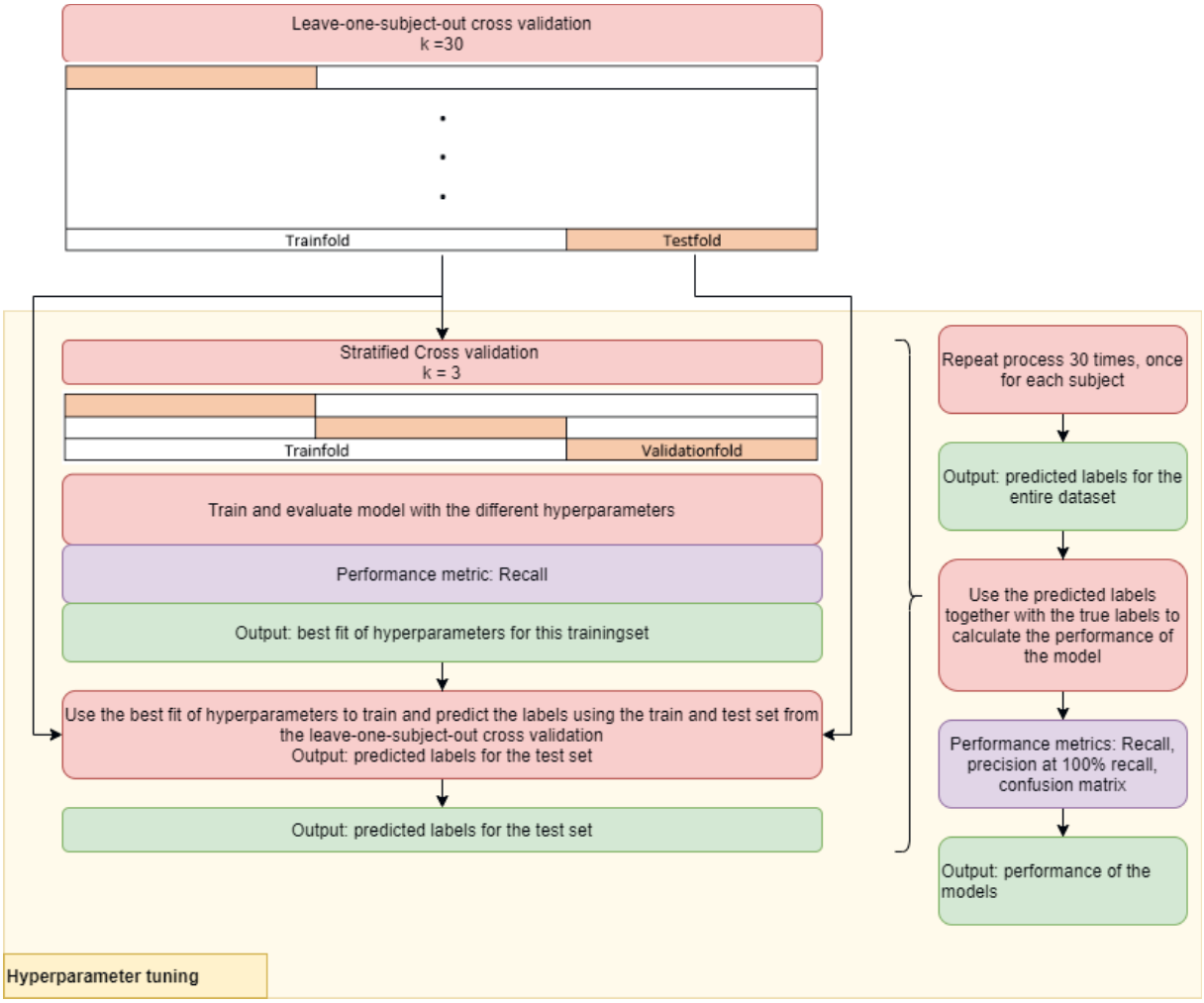


Figure 6.5: The pipeline of the hyperparameter tuning process.

Figure 6.5 presents the pipeline of this phase. The attempts from all subject minus one are used as trainingset to find the best hyperparameters for this subject. The trainingset is again divided by stratified cross validation into a trainingset and a validation set, to lower the bias and variance of the model. The model with the hyperparameters that resulted in the best recall is used to evaluate the performance of the model when classifying the unseen attempts from the testset. This process is repeated thirty times, once for every subject. This results in a complete list of predicted labels for all attempts, which is compared to the labels assigned to the attempts by the professional to determine the performance of the model.

The Long Short Term Memory (LSTM) model (see appendix A.5 for details),

which is a Recurrent Neural Network (RNN), is trained with the smoothed data as input, as this model takes into account the time-series component of the data. An $LSTM$ is used as the RNN as it is found that LSTMs outperform other RNNs when long time lags are involved in the task [62]. Additionally, LSTMs have the advantage that these do not suffer from the vanishing gradient problem; this happens when the gradients shrinks very fast becoming extremely small which does not contribute much to learning.

The LSTM model is optimized by evaluating one to ten hidden layers, with a step size of five, and one to 101 hidden neurons per layer, with a step size of fifty. If results show other values could improve the LSTM model, these values are used as well.

The Support Vector Machine (SVM) (explained in appendix A.8), boosted decision trees (explained in appendix A.7), and a Radial Basis Function Neural Network ($RBFNN$) (explained in appendix A.6), are trained using the different featuresets as input.

The SVM is evaluated with different kernel functions; a linear, Radial Basis Function (RBF), and a sigmoid kernel function.

The boosted decision trees is trained by applying two boosting algorithms; the real boosting algorithm, which uses the predicted class probabilities to update the weights, and the discrete boosting algorithm, which adapts the weights based on the errors in the predicted labels.

An $RBFNN$, which is a form of an Artificially Neural Network (ANN), is chosen to use as ANN in the present study as Manoharan et al. (2008) [60] found that an RBFNN gives better results compared to a normal back propagation neural network. The RBFNN is optimized by heuristically determining the K of the K-means clustering function (see appendix A) by using values between two and twenty, with a step size of two.

Phase 2: Balancing the dataset

When the performance of the models in the first phase show that balancing the dataset could improve the performance, the appropriate balancing technique is chosen, and the hyperparameter tuning process is repeated with the balanced dataset. The performance of the models when using the balanced dataset is compared to the performance using the imbalanced dataset. To determine the right balancing technique, the distribution of the errorclasses in combination with the confusion matrices from phase one are used. Examples of balancing methods are upsampling the minority class by randomly duplicating entries, or downsampling the majority class by randomly removing entries.

Phase 3: Stacking different models

Stacking is an ensemble method which combines the predictions of multiple single models into a prediction by a final estimator, to improve the predictions. A graphical visualization in which three single models are stacked, is shown in figure 6.6.

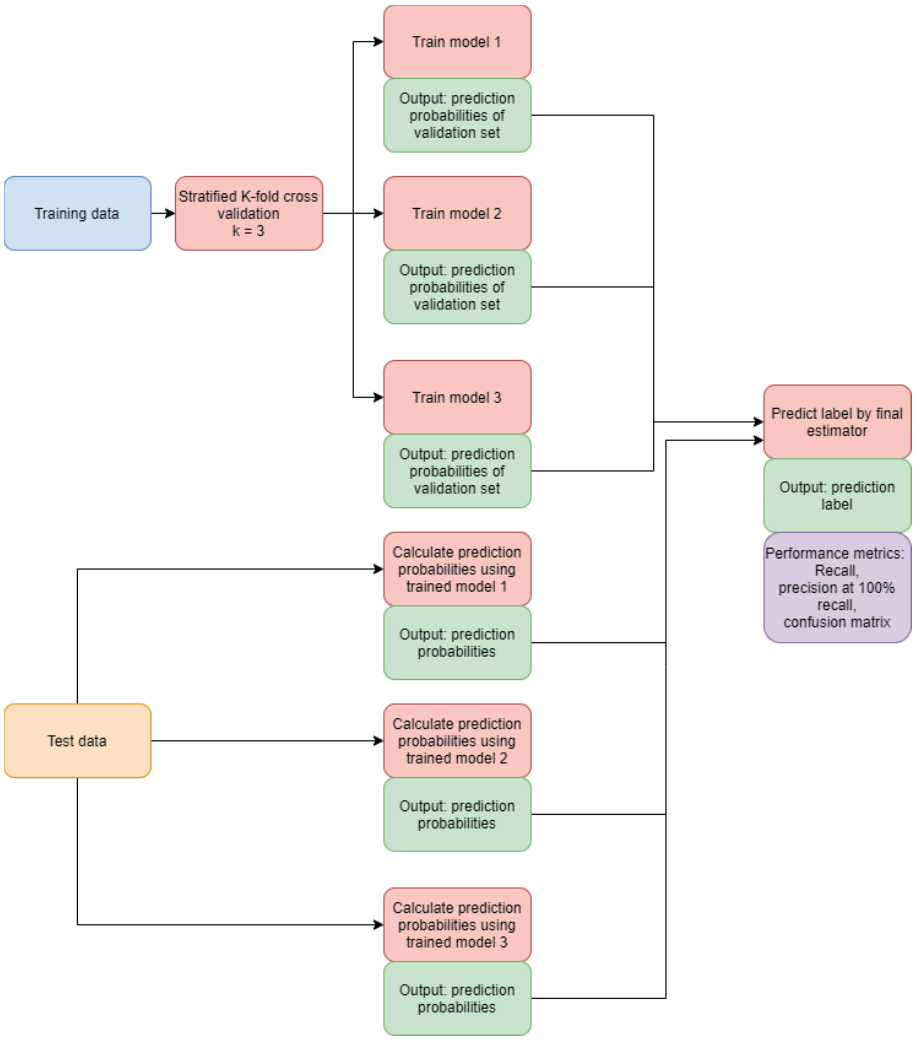


Figure 6.6: A graphical visualisation of model stacking with three single models.

The best performing models are stacked. The models are first divided in a validationset and trainingset by stratified cross-validation. This trainingset is used to train the single models using the featureset and hyperparameters it performed best with. For every attempt in the trainingset, the probability that the attempt is predicted as a certain label are calculated for the validationset by all single models and fed into a final estimator. This final estimator is evaluated using the testset to determine the performance of the stacked model. This process is repeated for every subject, using the leave-one-subject-out cross validation approach explained in section 6.1.3. The expectation is that combining the prediction by the different models, results in

a stronger model as information from multiple models is used. However, stacked models are prone to overfitting, meaning that the model performs very well on seen attempts, but badly on unseen attempts. Therefore, the performance of the stacked model based on the trainingset is evaluated as well.

6.1.6 Proposed decision tree

A decision tree consisting of three stages is proposed, presented in figure 6.7.

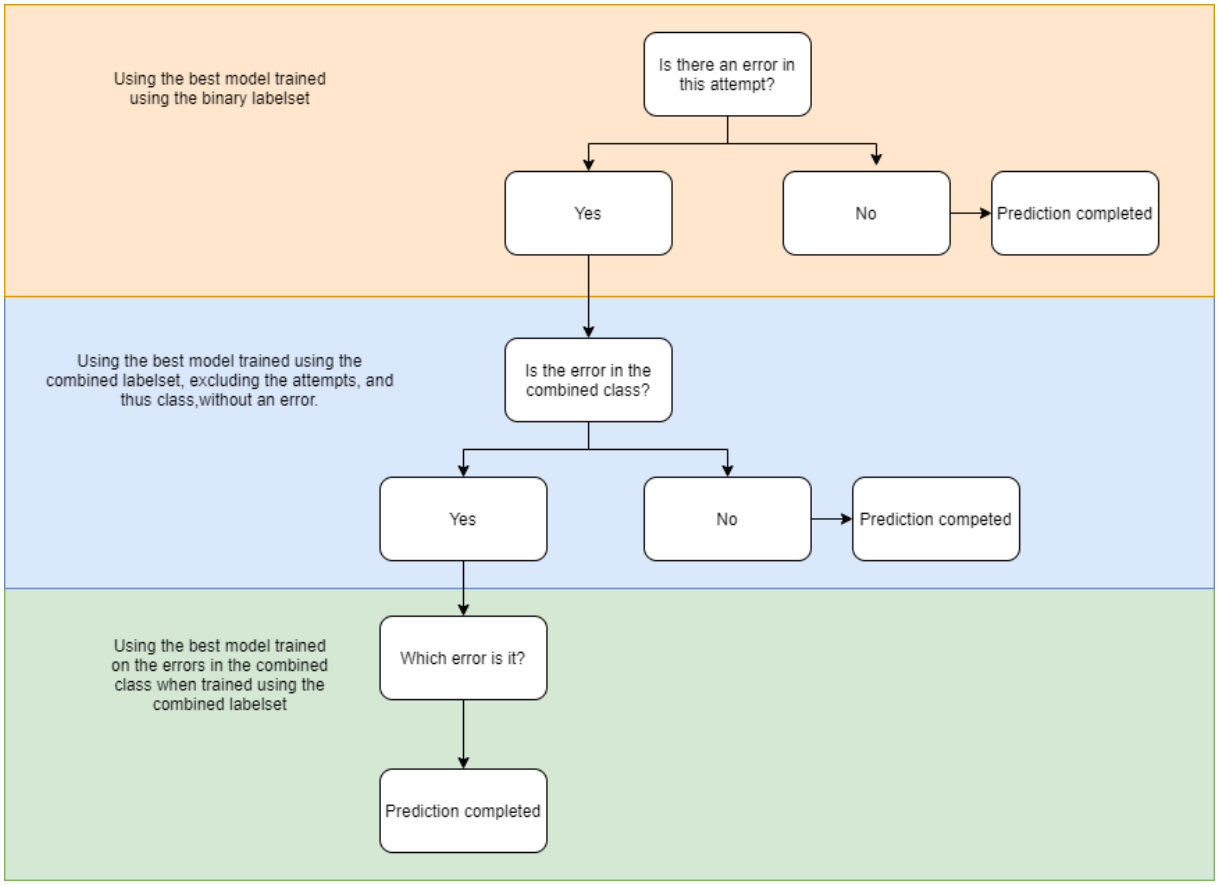


Figure 6.7: The proposed decision tree

The first stage predicts whether an attempt contains an error. If not, the process is finished, and this prediction is used as the outcome for this attempt. If the prediction shows that the attempt contains an error, it is processed by the second stage. This stage is a model trained on the dataset with the combined labelset, but without the class representing the attempts with no errors. If the model predicts the attempt to be in a class with a single error, the process is finished, and the prdicted label is used. However, if the attempt is predicted to be in the class with the combined errors, the attempt is processed by the third stage, which is a model based on the

errors in the combined class of the combined labelset, to predict which error is in the attempt.

To choose the best models for stage two and three, phase one to three of the model training process are executed for these labelsets.

If the prediction in stage one is wrong, the attempt cannot be predicted in stage two, as no trainingset is available. Therefore, in this case the prediction is a random guess from a sample of errors which could be predicted in this stage, taking the label distribution into account. This also holds for the third stage of the proposed decision tree.

6.2 Inter-annotation study

To determine the agreement of the error detection by professionals, the inter-rater agreement, which represents the agreement between professionals, and the intra-rater agreement, which represents the agreement of one professional over time, were calculated. The pipeline of this study is presented in figure 6.8. The different stages are elaborated in the remainder of this section.

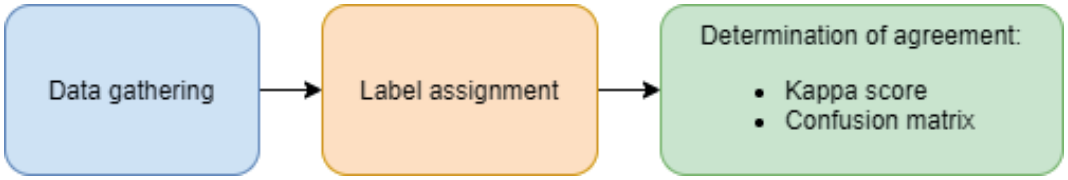


Figure 6.8: The pipeline of the inter-annotation study.

6.2.1 Data gathering

To gather the data for the inter-annotation study, thirteen adults between eight-teen and thirty-three years old were asked to perform two spirometry tests, each three to eight attempts; one with feedback from the researcher, and one with feedback from the metaphors of the SpiroPlay system to mimic the hospital experiments explained in section 6.1.1. The participants were instructed beforehand how to perform spirometry and their age and height were asked to set the GLI standards. The participants were informed by an information letter (refer to appendix D.2) and were asked to sign a consent form (refer to appendix D.3) when they agree to take part in the study and to be recorded.

As the data is privacy sensitive, the professionals who assessed the data were asked to sign a form as well in which they state that the data will be deleted by them

as soon the assessment is done, and that the data will not be used for other goals than the error detection for this study. The form can be found in appendix D.4.

The attempts are typically evaluated by looking at the person during the attempt, and by evaluating the resulting flow-volume curve. As it is not possible for the professionals to be present during the measurements, video-recorded attempts will be provided along with the corresponding curves. The participants were recorded from two angels; the whole body from the side, and the face from an oblique side view. A test recording was performed and analyzed by two professionals to assess if the camera positions were correct. On the basis of the feedback on this recording, the camera recording the face was changed from a front view to an oblique side view.

The recordings of both cameras and the curve of the attempt were synchronized to presented them in one video, all at the same time. This made it more accessible for the professionals to label the data.

As the experiment took place during the COVID-19 crisis, the participants and the researcher wore latex gloves, which were renewed after every participant, and the spirometer, tablet and nose clip were desinfected between participants with alcohol. A very simplified version of the setup is shown in figure 6.9 to represent the angles of the cameras.

Figure 6.9: A simplified version of the study setup, showing the angles of the cameras.

6.2.2 Label assignment

To be able to calculate the inter-rater agreement, three professionals independently labeled attempts by assessing which error(s) they detect in the attempts. More than

one label could be given, as more than one error can occur during an attempt. Refer to table 6.1 for an overview of the different errors. To be able to calculate the intra-rater agreement, the professionals were also asked to label the same data twice with two weeks in between.

The labelsets used to find the best error detection algorithm were mimicked to be able to compare the inter-rater agreement to the performance of the error detection algorithm. This means that three labelsets were created; a binary labelset in which the attempts were divided in two classes representing the attempts with and without errors, a combined labelset in which the attempts labeled as an error from the error-classes in the 'other: namely' category were combined, and a labelset in which all attempts preserved the label given by the professional.

6.2.3 Determination of the agreement

To calculate the inter- and intra-rater agreement, Cohen's kappa score (refer to appendix A.1 for the calculation) is calculated, which is one of the leading approaches in calculating inter- and intra-rater agreement. [63].

As explained in section 6.2.1, the professionals label the data twice with two weeks in between. The inter-rater agreement is calculated for all combinations of rounds between the professionals, to get a complete overview.

The intra-rater agreement is calculated for every professional, for every labelset.

The interpretation of the Cohen's kappa score is presented in table 6.5.

Cohen's kappa score	**Interpretation**
0 - 0.20	None
0.21 - 0.39	Minimal
0.40 - 0.59	Weak
0.60 - 0.79	Moderate
0.80 - 0.90	Strong
$>$0.90	Almost perfect

Table 6.5: Interpretation of the Cohen's kappa score [64].

Next to the kappa score, confusion matrices were created to assess which errorclasses are confused between the professionals, and by one professional over time.

6.3 Comparison of coaching by a metaphor versus by a professional

The proposed system includes metaphors which are used to steer the blowing behaviour of the user. The goal of this part of this research is to determine if there is a significant difference in blowing behaviour when the participant is coached by the metaphors compared to when coached by the professional, to evaluate if the metaphors are a good way of coaching asthma patients during (home) monitoring.

The process from the data gathering to the analysis consists of several steps which are visualized in figure 6.10. The blue boxes denote a preparing step, and the orange box represents the analysis part of the pipeline.

Figure 6.10: The pipeline of the comparison between coaching by the metaphors and the professional.

6.3.1 Data gathering

During the hospital experiments, explained in section 6.1.1, thirty children were asked to perform two tests, each three to eight attempts; one with a metaphor and one without. During one test, they were coached by a trained professional, while during the other test, they were coached by the metaphors. Half of the children was first coached by a professional, followed by a test coached by a metaphor, while the other half performed the tests in the reverse order. This data is used to compare the coaching by the metaphors and by the professional.

6.3.2 Data preprocessing

To reduce noise, the data is cleaned to only maintain the attempts without an error for the comparison of the FVC, FEV_1, and PEF values. If subjects did not blow attempts without errors, these subjects are not taken into account in the comparison.

For the comparison of the number of errors produced when coached by the two approaches, all attempts, except for the outliers and attempts with label 'uncertain', are taken into account.

6.3.3 Data analysis

The mean values of the PEF, FEV_1, and FVC of the participants coached by the metaphors and the mean values of these parameters when coached by the professional were calculated to compare the two situations. Besides, the number of errors produced per coaching situation and per subject were calculated.

The first step in the comparison of every parameter and the number of errors is to determine if the dependent variable is normally distributed. This was examined by a QQ-plot and by performing a Shapiro-Wilk test. The threshold for this test to not reject the H0 hypothesis, which states that the dependent variable is normally distributed, was set at an alpha value of 0.05, meaning that p-values above 0.05 show that the H0 hypothesis stands.

If the dependent variable was not normally distributed, a Wilcoxon test was performed to assess if there is a statistically significant difference in the PEF, FVC, and FEV_1 values, and the number of errors between the two coaching strategies. Otherwise, a paired sample T-test was performed. The threshold to not reject the H0 hypothesis, this hypothesis states that there is no significant difference, was set at an alpha value of 0.05.

The three tests and the QQ-plot are described in appendix A.10.

Chapter 7

Results

7.1 Error detection

An error detection algorithm for detecting errors in spirometry attempts using machine learning techniques is designed and evaluated. The data from the hospital experiments was used as input. Section 7.1.6 describes what the effect is on the performance of the models when extending the dataset with the data from the inter-annotation study.

7.1.1 Dataset

The data is gathered during hospital experiments, explained in section 6.1.1. During this study, the children performed two spirometry tests, each existing of three to eight attempts. One test is performed with feedback from a professional which supervised the test, during the other test the child is coached by metaphors, described in section 2.4. The dataset used to train and evaluate the models on are the attempts from both tests of all children.

A professional labeled the attempts during these tests by viewing the curves and the children performing the tests. She used the form presented in appendix D.1 to file the errors she detected.

The 'other, namely:' case was used extensively, resulting in more errorclasses than given in table 6.1. The complete list of errorclasses with their codes is presented in table 7.1.

7.1.2 Preprocessing

The data was prepocessed before used during model training. This included cleaning, exploring, and filtering the data. Also, different labelsets and featuresets were generated.

Error	**Code**
Extrapolated volume $<$5% of FVC or less than 0.15 L at the start of the expiration	1
Obstructed mouthpiece	2
An extra breath taken during the attempt	3
Flow leak	4
Duration of $<$3 s, a plateau in the VT curve, or if the person can/should continue exhaling	5
No maximal effort	6
Cough	7
Wrong posture	8
Other, namely: No maximal inhalation	9
Other, namely: Growl	10
Other, namely: Waiting too long between inhalation and exhalation	11
Other, namely: Started too early with exhalation	12
Other, namely: Exhalation, before inhalation, in apparatus	13
Other, namely: Variable effort	14
Other, namely: Making a sound during exhalation	15
Other, namely: No peak, but flat top in the flow-volume curve	16
Other, namely: No inhalation in the apparatus	17
Other, namely: Sawtooth curve (tech error)	18
Other, namely: Wrong order of inhalation and exhalation	19
Other, namely: First exhaling softly, and then with more power	20
Other, namely: The tube is not far enough in the mouth and the lips are not tight enough around the tube	22
No error	0
Uncertain	-1

Table 7.1: Complete table of error codes

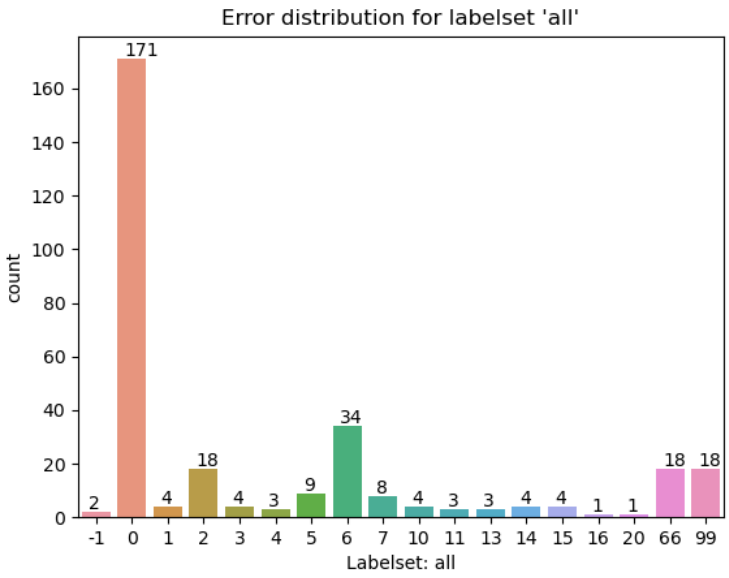


Figure 7.1: Histogram representing the distribution of the labeled errors, including the attempts labeled as uncertain (-1) and the outliers (99).

Cleaning the data

As can be seen in the histogram (Figure 7.1), two attempts with label -1, which represent the label 'uncertain', were present in the data. As these serve no purpose in training, testing, or validation, these were removed from the dataset. During some attempts, the system used did not measure the attempt well. These curves resulted in sawtooth curves, and were neglected by the practioner. As this is an error from the system which is expected to be repaired before used in real life, these attempts were removed from the dataset.

Several attempts were labeled with more than one errorlabel. As the data from this attempt will bring noise into the dataset when this attempt is labeled as one of the errors present in the attempt, the attempts with more than one error were combined into one class to which label 66 was assigned.

As can be seen in the first figure of 7.2, there is a lot of noise in this data; a second peak, which is a rough copy of the first one, and the curve is rugged. Therefore, the data was filtered and smoothed to only keep the useful information. The parts of the data before the inhalation and after the exhalation were automatically removed by a self-written algorithm. Additionally, the curve was smoothed to reduce noise. The process of filtering and smoothing is shown in figure 7.2.

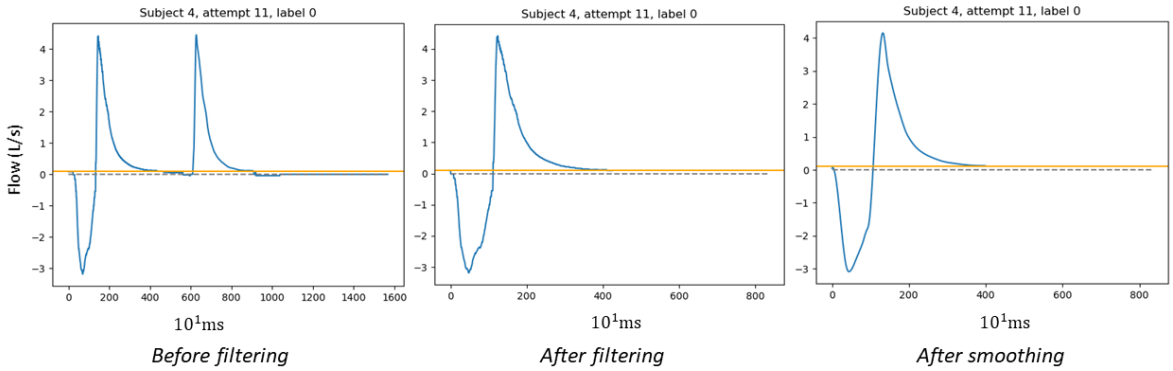


Figure 7.2: The process of filtering and smoothing.

Data exploration

As can be seen in figure 7.3, the data is very imbalanced with a ratio of approximately 5:1 between the largest and second largest class. Besides, most error classes have a small number of datapoints.

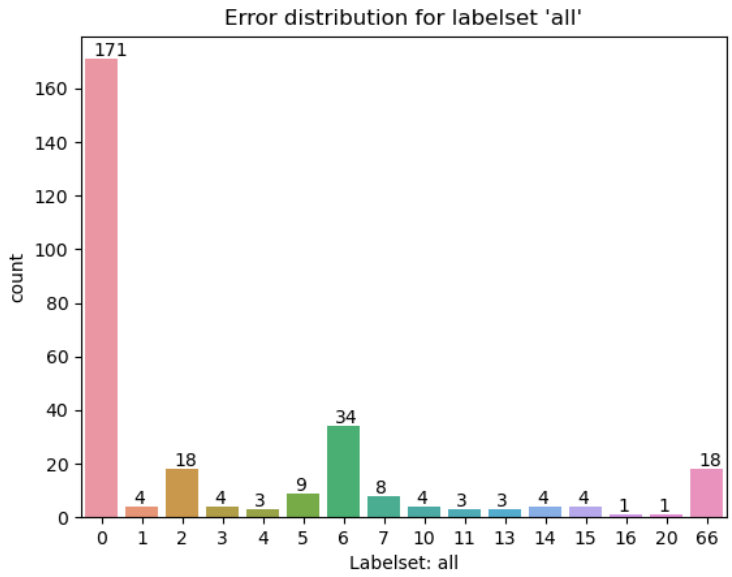


Figure 7.3: Histogram representing the distribution of the labeled errors, excluding the attempts labeled as 'uncertain' (-1) and the outliers (99).

To show the extend to which the different errorclasses are linearly related, a correlation plot, presented in figure 7.4, was created. One random sample from each errorclass was taken and the correlations were calculated. To give a fair overview, this process was repeated a hundred times, after which the correlations of these processes were averaged. The plot shows that the attempts from the different error classes are heavily related.

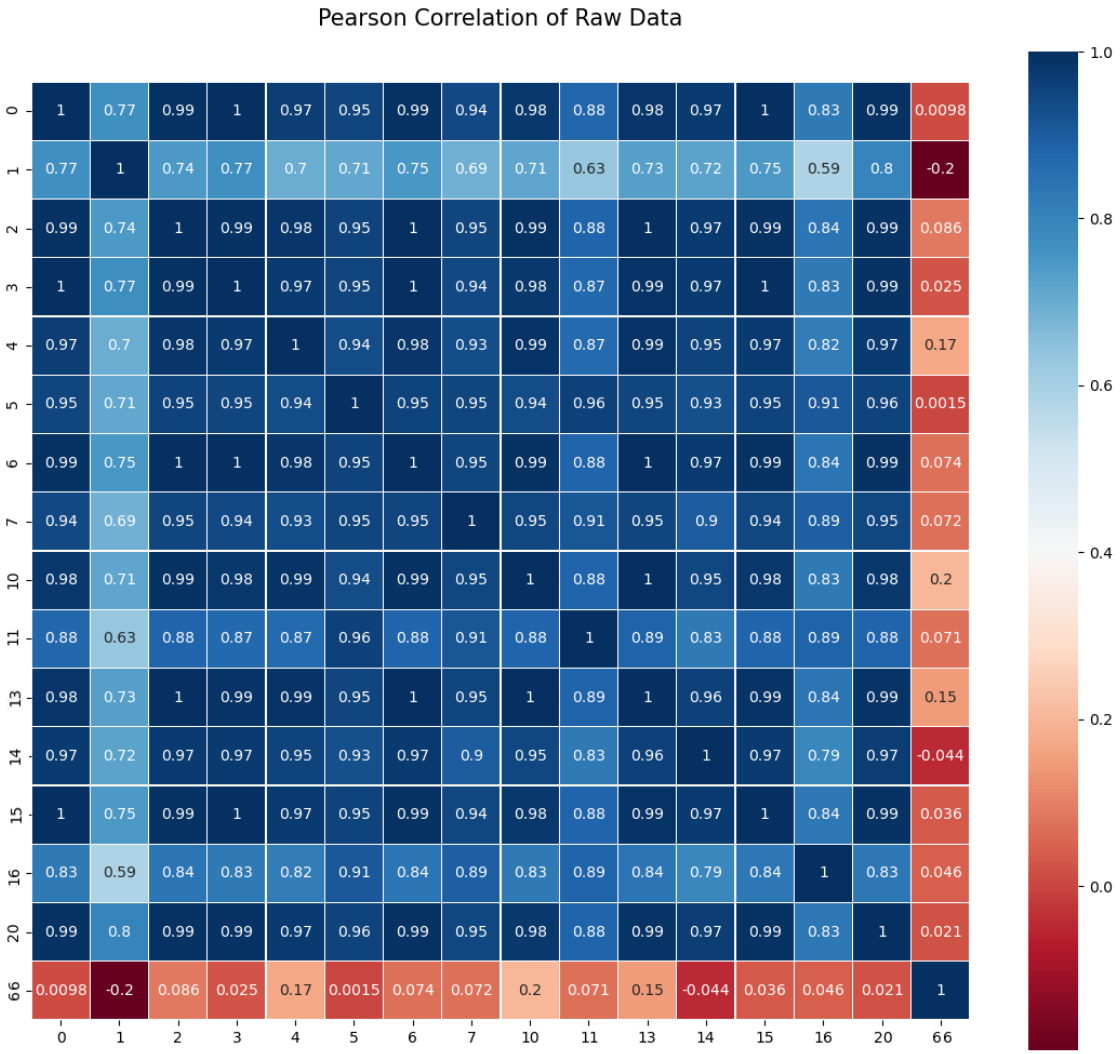


Figure 7.4: A correlation plot showing the extend to which the different errorclasses are linearly related.

Outlier removal

Fourteen datapoints were labeled as outliers as they were more than 2.5 standard deviations from the mean value of the FVC or the FEV_1 values from this person. Besides, the curves were assessed visually to find outliers. On this basis, six datapoints were labeled as outliers. These attempts crossed the zero line before exhalation started, and thus were not representable. Some outliers from both assessment techniques overlapped, which resulted in a total amount of eight-teen outliers. The outliers were attempts with labels two (obstructed mouthpiece), five (terminated too early), seven (cough), ten (growl), eleven (waiting too long between inhalation and exhalation), twelve (started too early with the exhalation), thirteen (exhalation before inhalation in the apparatus), fifteen (sound during exhalation), seventeen (no inhalation in apparatus), and sixty-six (multiple errors).

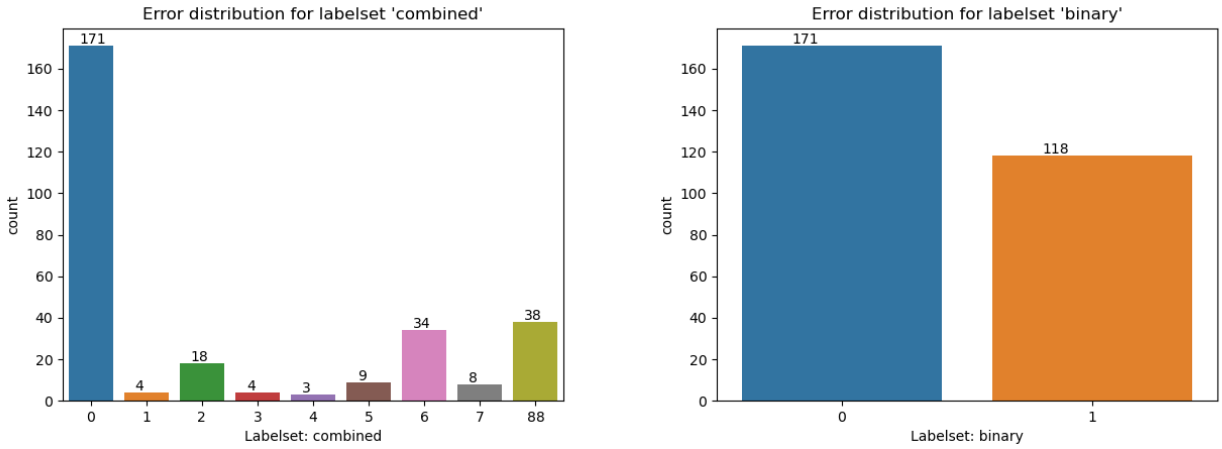
Generating different labelsets

As explained in section 6.1.2, three labelsets were created. The first labelset consists of two classes; one class consisting of all attempts containing an error, and one class with all technically correct attempts. The second labelset consists of nine classes. These are the single classes with labels zero to eight, as these are based on the criteria to be met according to Miller et al. (2005) [3], and are important to provide specific feedback about. The combined class contains all attempts labeled as an attempts from one of the 'other, namely:' classes. These are the errorclasses ten to twenty. When an error from these categories is produced, it is important to feed back that an error is made. However, as these errors are not as important as the errors from classes one to seven according to Miller et al. (2005) [3], general feedback is sufficient. The attempts containing multiple errors (labeled with sixty-six) are also included in the combined class, as it is also not possible to provide specific feedback based on these attempts. The combined class is assigned label eighty-eight to distinguish this class easily from the single errorclasses. The third labelset is a labelset in which every single class keeps its label, to evaluate if it is feasible to distinguish all errorclasses.

For the remainder of this report, the labelsets will be referred to as 'binary' , 'combined', and 'all'. The histograms of the binary, combined, and all labelset are shown in figures 7.5b, 7.5a, and 7.3. The labels for each error class are presented in table 7.2.

Original error label	All	Combined	Binary
0	0	0	0
1	1	1	1
2	2	2	1
3	3	3	1
4	4	4	1
5	5	5	1
6	6	6	1
7	7	7	1
10	10	88	1
11	11	88	1
13	13	88	1
14	14	88	1
15	15	88	1
16	16	88	1
20	20	88	1
66	66	88	1

Table 7.2: The error labels for the different classes in the three labelset.



(a) The combined labelset

(b) The binary labelset

Figure 7.5: Histograms showing the labeled error distribution of the combined and binary labelset. The classes with labels 10 to 20 are combined in the combined class 88 in the combined labelset. The binary labelset consists of two classes; one with all technically correct attempts, and one with all attempts containing an error.

7.1.3 Model training

After the data was preprocessed, models were trained in three phases to find the best fit. These phases are hyperparameter tuning, balancing, and stacking. After this section, the proposed decision tree is evaluated.

Phase 1: Hyperparameter tuning

As explained in section 6.1.5, during the first phase four different models were trained with different hyperparameters. Also, three labelsets were used refered to as ' all' , ' combined' , and ' binary' , as explained in section 7.1.2. Besides, five different featuresets extracted from the dataset were used, as explained in section 6.1.2.

During hyperparameter tuning, it was found that the LSTM always resulted in the best performance when adopting one node and one layer, when the hyperparameters one to 101 nodes, with a stepsize of 50, and one to eleven layers, with a step size of five, were evaluated. Therefore, the possible hyperparameters were changed to one to five hidden layers and hidden nodes, with a step size of two, to evaluate if models with layers and hidden nodes in this range performed better than a model with one layer and one hidden node.

The performance represented by the precision, recall, F1-score, and the precision at 100% recall are shown in table C.1, C.2, and C.3 in the appendix. The recall and precision at 100% recall are presented in table 7.3 for all labelsets, as these are the primary and secondary performance metrics.

This table shows that the best performance based on the recall is met using the SVM, using the filtered features as input for the binary labelset (0.737), and the spirometry parameters for the labelsets 'combined' (0.356) and 'all' (0.305). When using the precision at 100% recall as metric, the best models are the SVM using the spirometry parameters for labelsets 'binary' (0.605) and 'combined' (0.125), and the RBFNN using the spirometry parameters as input and labelset 'all' (0.065).

When looking at the results of the binary labelset, one can see that the SVM outperforms the other models, independently of the featureset. Additionally, the boosted decision trees outperforms the RBFNN based on recall. The LSTM outperforms the RBFNN using the unfiltered featureset, and the boosted decision trees using the unfiltered featureset in combination with the spirometry parameters. When looking at the precision at 100% recall, the difference is in the RBFNN outperforming the boosted decision trees when using the spirometry parameters as input. The LSTM outperforms or performs equal to all RBFNN, except for the RBFNN using the spirometry paremeters as input.

Featureset	Model	Binary		Combined		All	
		Recall	***Precision at 100% recall***	***Recall***	***Precision at 100% recall***	***Recall***	***Precision at 100% recall***
Spirometry parameters	RBFNN	0.525	0.539	0.203	0.117	0.178	0.065
	Boosted decision trees	0.653	0.532	0.229	0.113	0.119	0.063
	SVM	0.720	0.605	0.356	0.125	0.305	0.063
Filtered features	RBFNN	0.559	0.500	0.178	0.111	0.144	0.063
	Boosted decision trees	0.610	0.541	0.136	0.111	0.059	0.063
	SVM	0.737	0.573	0.263	0.111	0.195	0.063
Filtered features + spirometry parameters	RBFNN	0.500	0.500	0.127	0.111	0.161	0.063
	Boosted decision trees	0.559	0.504	0.169	0.111	0.144	0.063
	SVM	0.720	0.557	0.229	0.115	0.195	0.063
Unfiltered features	RBFNN	0.381	0.500	0.017	0.113	0.025	0.063
	Boosted decision trees	0.475	0.509	0.017	0.111	0.025	0.063
	SVM	0.508	0.518	0.051	0.111	0	0.063
Unfiltered features + spirometry parameters	RBFNN	0.466	0.500	0.042	0.111	0.017	0.063
	Boosted decision trees	0.449	0.504	0.059	0.111	0.025	0.063
	SVM	0.500	0.511	0.059	0.111	0	0.063
Smoothed data	LSTM	0.449	0.500	0.102	0.112	0.127	0.063

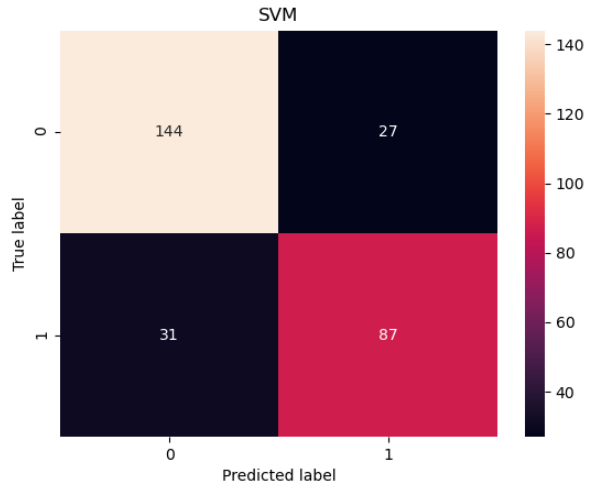
Table 7.3: The performance of the models after hyperparameter tuning using the different labelsets, represented by the recall and precision at 100% recall.

For the combined labelset, the SVM outperforms the boosted decision tree, and the RBFNN, based on recall for all featuresets, except for the unfiltered featureset in combination with the spirometry parameters, where the SVM performs equally to the boosted decision tree. Secondly, the boosted decision trees outperforms the RBFNN for all featuresets, except the filtered features, where the RBNN outperforms, and the unfiltered features, where the performance is equal. The LSTM outperforms all models using the two unfiltered featuresets as input. When using the precision at 100% recall as performance metric, the SVM outperforms or performs equal the the other models, except when using the unfiltered features as input as the RBFNN outperforms the other models. The LSTM is outperformed by the models using the spirometry parameters, the SVM using the filtered features in combination with the spirometry parameters, and the RBFNN using the unfiltered features as input.

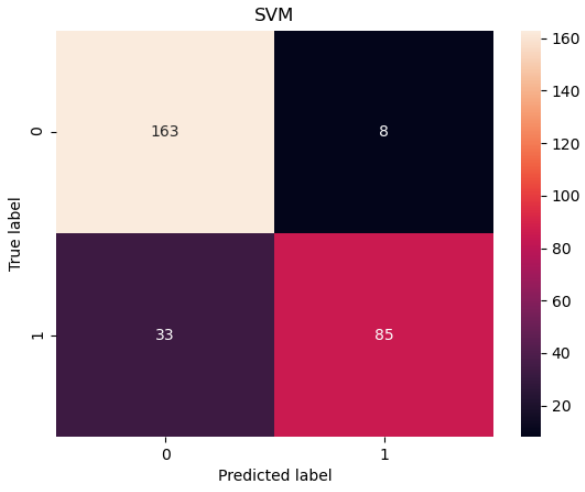
When looking at the recall for labelset 'all', we see that the SVM outperforms the other models using all featuresets, except the two unfiltered featuresets where the recall of the SVM is zero. For these featuresets, the boosted decision trees outperforms or performs equal to the RBFNN. For the other featuresets, the RBFNN outperforms the boosted decision trees. The LSTM outperforms all models using the two unfiltered featuresets, and the boosted decision trees when using the spirometry parameters or the filtered featureset as inputs. The precision at 100% recall is equal for all models, except for the RBFNN using the spirometry parameters, which precision at 100% recall is 0.002 higher.

When comparing the labelsets, we see that the performance decreases with the increase in the number of errorclasses.

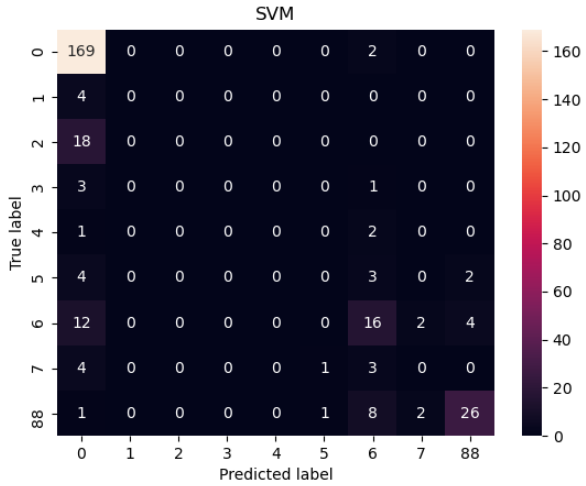
The confusion matrices of the best models based on the recall and precision at 100% recall per labelset are shown in figure 7.6.



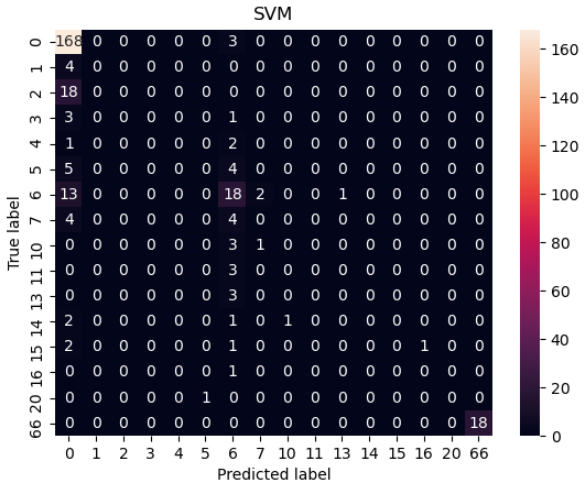
(a) Binary: recall



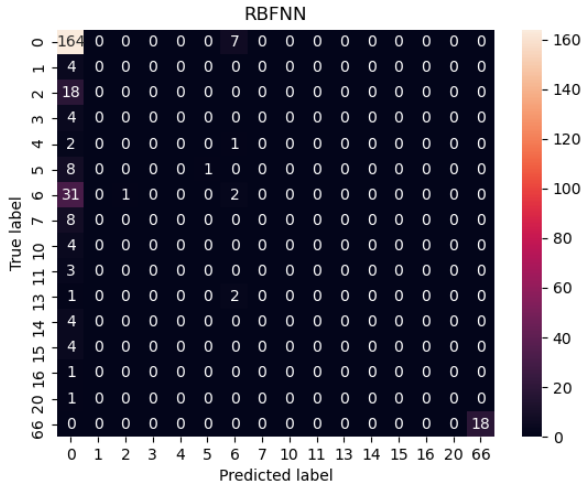
(b) Binary: precision at 100% recall



(c) Combined



(d) All: recall



(e) All: precision at 100% recall

Figure 7.6: The confusion matrices of the best models after hyperparameter tuning, based on the recall and the precision at 100% recall.

These matrices show that a lot of datapoints are unfairly labeled as zero, which label represents the technically correct attempts. From the histogram in figure 7.3, we see that the data is skewed towards this errorclass. Therefore, the data was balanced to evaluate if this improves the performance of the models.

Phase 2: Balancing the dataset

The data is balanced by upsampling the minority errorclasses to contain an equal amount of datapoints as the majority class. Two ways of upsampling are evaluated; Random OverSampling (ROS) and Synthetic Minority Oversampling Tech-

nique (SMOTE). ROS simply duplicates random datapoints from the class to be upsampled. SMOTE selects two datapoints at random from one class which are near each other in the feature space. A line is drawn between those points and a new sample is created along this line. The advantage of SMOTE over simple oversampling is that more information is created. However, the downside is that the other classes are not taken into account, which can lead to datapoints overlapping with other classes. This is not the case with random oversampling.

Tables C.4 to C.9 in the Appendix present the performance of all models. For the binary labelset, the recall of fourteen models improved when using SMOTE, with a mean improvement over all models of 0.058. When using ROS, the recall of fifteen models improved, and one stayed the same. The mean improvement was 0.070. The precision at 100% recall improved when using SMOTE for six models, stayed the same for three models, and decreased for seven models, with a mean decrease of 0.00056. When applying ROS, eleven models improved, one stayed the same, and four decreased in performance. The mean increase in precision at 100% recall was 0.015.

For the combined labelset, all models except two improved in performance based on recall, with a mean improvement of 0.039 and 0.069 when applying SMOTE and ROS respectively. When looking at the precision at 100% recall, nine models improved when using SMOTE, with a mean improvement of 0.001. When using ROS, twelve models improved, with a mean improvement of 0.002.

When interpreting the tables of labelset 'all', we see that, based on recall, eleven models improved when using SMOTE and ROS, and the performance of one model stayed the same when using ROS. The mean improvements after applying SMOTE and ROS respectively were 0.035 and 0.040. When looking at the precision at 100% recall after applying SMOTE, two model improved, two models decreased in performance, and twelve stayed the same, with a very small mean decrease of $3.757 * 10^{-6}$. After applying ROS, the performance of three models increased, of one decreased, and of twelve stayed the same, with a very small mean improvement of $6.5 * 10^{-5}$.

Overall, an improvement in recall did not necessarily mean an improvement in precision at 100% recall. Also, the classifiers that improved in performance were not the same for the different featuresets, labelsets, and balancing techniques. Thirdly, the performance gain was bigger when looking at recall, than when comparing the precision at 100% recall.

Table 7.4 shows the performance of the best models of every classifier with the featureset and balancing technique they performed best with, as these models are used in the stacking phase. The decision of best models was based on the recall, as the only models outperforming these models based on precision at 100% recall were the RBFNN and LSTM for labelset 'binary'. The recall and precision at

Labelset	**Model**	**Featureset**	**Balancing technique**	**Recall**	**Precision at 100% recall**
Binary	RBFNN	Filtered features + spirometry parameters	ROS	0.644	0.518
	Boosted decision trees	Spirometry parameters	ROS	0.686	0.596
	SVM	Filtered features	ROS	0.864	0.678
	LSTM	Smoothed data	ROS	0.559	0.500
Combined	RBFNN	Spirometry parameters	ROS	0.288	0.122
	Boosted decision trees	Spirometry parameters	ROS	0.305	0.120
	SVM	Spirometry parameters	ROS	0.525	0.134
	LSTM	Smoothed data	ROS	0.153	0.113
All	RBFNN	Spirometry parameters	ROS	0.220	0.065
	Boosted decision trees	Filtered features	SMOTE	0.263	0.063
	SVM	Spirometry parameters	ROS	0.322	0.063
	LSTM	Smoothed data	None	0.127	0.063

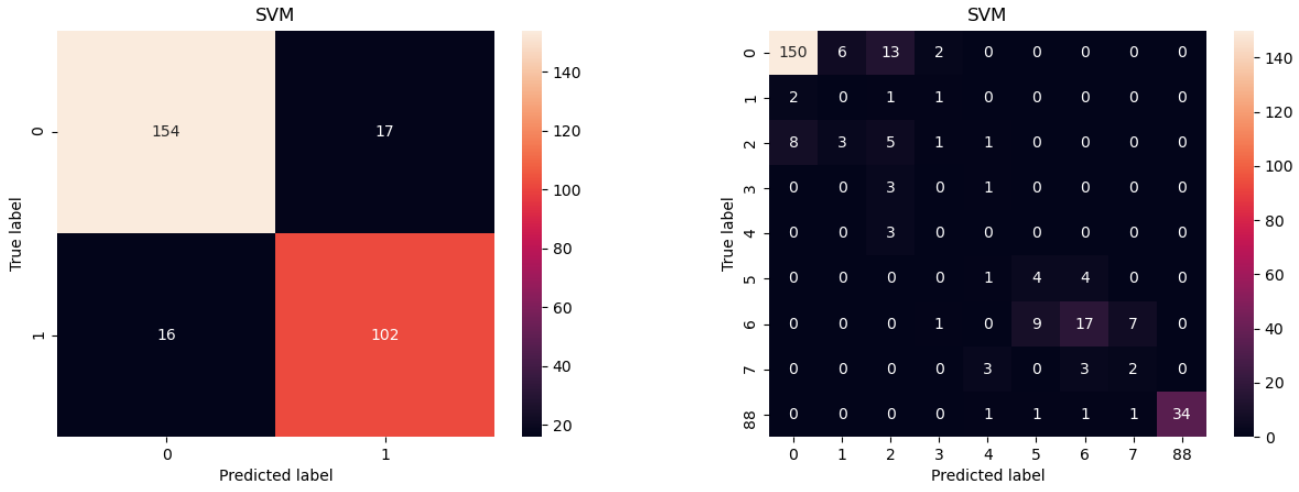
Table 7.4: The best models of every classifier per labelset after hyperparameter tuning and balancing. The table shows the featureset and balancing technique the models perform best with.

100% recall of the RBFNN using the spirometry parameters, and balancing technique ROS, was 0.559 and 0.541. As the decrease in recall, compared to the best performing RBFNN based on recall, was 3.570 as big as the increase in precision at 100% recall, the decision was made to use the RBFNN with the best recall as best performing RBFNN. The LSTM performed better when looking at precision at 100% recall after applying SMOTE instead of ROS, with a recall of 0.441, and a precision at 100% recall of 0.509. The decrease in recall was 13.182 as big as the improvement in precision at 100% recall, and thus the LSTM using ROS was used as best performing LSTM.

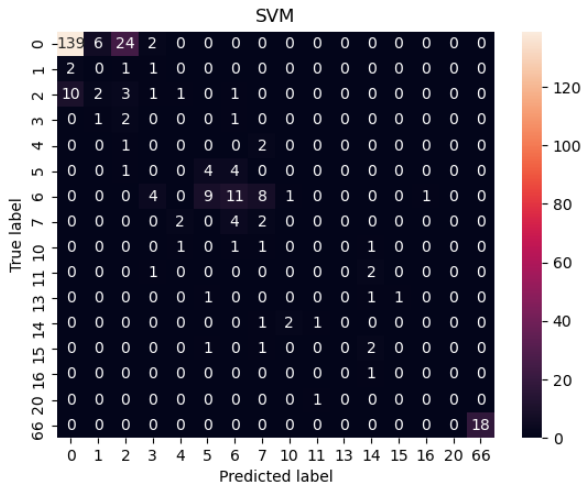
The best performing models based on recall are the SVMs for all labelsets. However, when looking at precision at 100% recall, the RBFNN outperforms the SVM for labelset 'all'. As the decrease in recall is 34.514 as big as the increase in precision at 100% recall, the SVM is chosen as the best model for this labelset.

The increase in performance after balancing for the binary labelset, comparing the best model based on recall for the increase in recall, and the best model based on precision at 100% recall for the increase in performance based on this metric, we see an improvement of respectively 0.127 and 0.073 for the recall and precision at 100% recall. The improvement in performance for the combined labelset is 0.169 and 0.009 for respectively the recall and the precision at 100% recall. The improvement for labelset 'all' is 0.017 and 0 for the recall and precision at 100% recall. However, as the decision was made to use the SVM as best performing model, there is a decrease of 0.003 when comparing this model to the best model based on precision at 100% recall before balancing.

The confusion matrices of the best models are shown in figure 7.7.



(a) Binary *(b)* Combined



(c) All

Figure 7.7: The confusion matrices of the best performing models for all labelsets, after hyperparameter tuning and balancing. The best models are the SVMs, using the filtered features as input for labelset 'binary', and the spirometry parameters for labelsets 'combined' and 'all'.

When interpreting the confusion matrices, we see that less attempts are confused with the zero errorclass (technically correct attempts) compared to the confusion matrices of the models before balancing. However, for the labelsets 'combined' and 'all', attempts from errorclasses one (unsatisfactory start) and two (obstructed mouthpiece), are still mostly confused with the zero errorclass (technically correct attempts).

For the combined labelset, five attempts from errorclass two (obstructed mouthpiece), four attempts from errorclass five (terminated too early), and two attempts from errorclass seven (cough) are now classified right, while none of the attempts of these classes were classified right before balancing. Respectively one and eight more attempts from errorclasses six (no maximal effort), and eighty-eight (the combined class) were classified right. Errorclass six (no maximal effort) was confused with the zero errorclass (technically correct attempts) before balancing, while not anymore after balancing. Errorclass eighty-eight (combined errorclass) was more confused with errorclass six (no maximal effort), zero (technically correct attempts), and seven (cough) before balancing. The errorclass confused with the most different classes are errorclasses two (obstructed mouthpiece) and eigthy-eight (combined errorclass), which are both confused with four other classes. However, thirty-four out of thirty-eight of errorclass eighty-eight (combined errorclass) are classified right, while only five out of eightteen of errorclass two (obstructed mouthpiece were classified right. Also, the errorclass the other errorclasses are mostly confused with is

also errorclass two (obstructed mouthpiece).

When looking at the confusion matrix of labelset 'all', we see that three attempts of errorclass two (obstructed mouthpiece), four attempts of errorclass five (terminated too early), and two attempts of errorclass seven (cough) were classified right, while none of the attempts of these classes were classified right by the best models before balancing. Seven attempts less of errorclass six (no maximal effort) were classified right after balancing, which is now more confused with errorclasses three (an extra breath taken during the attempt), five (terminated too early) and seven (cough). Errorclasses that are confused with the most classes are errorclasses two (obstructed mouthpiece) and six (no maximal effort), which are both confused with five other classes. These are also the classes the other classes are confused most with.

Phase 3: Stacking

In this phase the probability predictions by the best models are combined and fed into a final model. The graphical visualization of this process can be found in figure 6.6. Models of different classifiers are used, as these are trained differently and thus create the most information for the final model when combined.

The models which were stacked are the models presented in table 7.4. However, as the LSTM performs much worse than the other models, this classifier was excluded.

A decision tree is used as the final estimator, as this classifier performs well on imbalanced data [65].

The correlation between the single classifiers, representing the information overlap, is shown in table 7.5, and is negligible to moderate [66].

	RBFNN	**Boosted decision trees**
Boosted decision trees	Binary: 0.568 Combined: 0.234 All: 0.239	
SVM	Binary: 0.583 Combined: 0.416 All: 0.461	Binary: 0.563 Combined: 0.434 All: 0.305

Table 7.5: The Pearson correlation between the single models used in the stacking for the different labelsets.

To evaluate if the stacked models overfitted, the performance of the models on the trainingset are shown in table 7.6. This shows that this performance is extremely high.

Labelset	Precision	Recall	F1-score	Precision at 100% recall
Binary	1.000	0.998	0.999	0.998
Combined	0.875	0.852	0.860	0.624
All	0.878	0.868	0.871	0.626

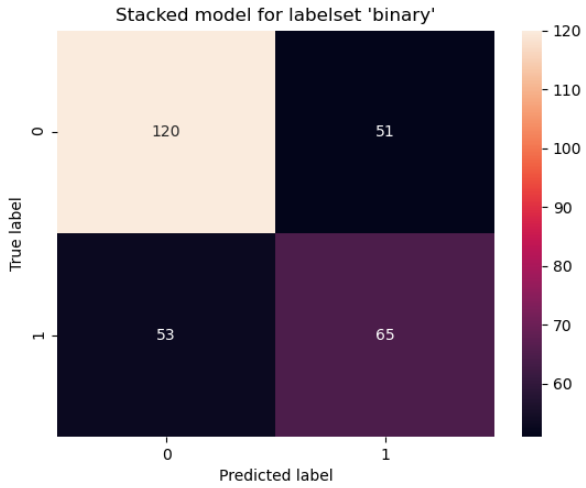
Table 7.6: The performance of the stacked models evaluated using the trainingset.

The performance of the stacked models when classifying unseen attempts can be found in table 7.7, and shows that the performance of the stacked models is lower than the performance of the single models.

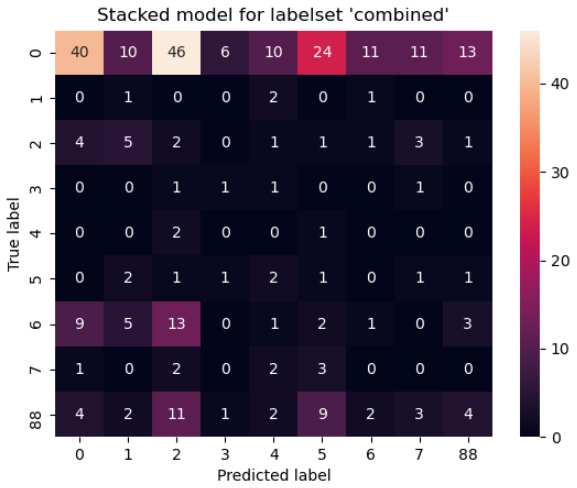
Labelset	Precision	Recall	F1-score	Precision at 100% recall
Binary	0.560	0.551	0.556	0.500
Combined	0.087	0.085	0.071	0.111
All	0.059	0.068	0.048	0.063

Table 7.7: The performance of the stacked models for the different labelsets

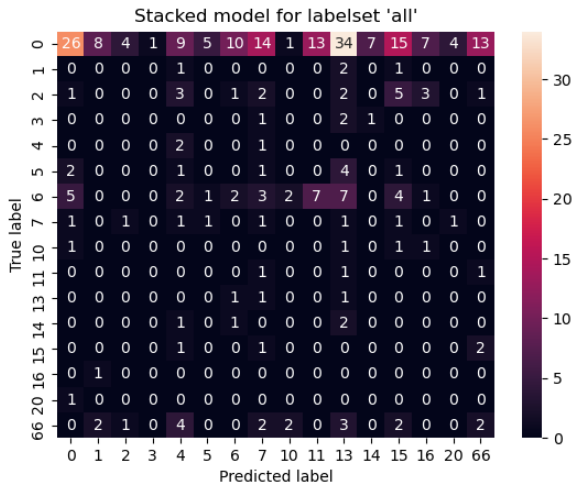
The confusion matrices, shown in figure 7.8, show that the errorclasses are more confused with the zero errorclass (technically correct attempts) than the best performing single models. Also, errorclasses six (no maximal effort), eighty-eight (combined errorclass) in the combined labelset, and sixty-six (multiple errors) when using labelset 'all' were classified well by the best single model, however poor by the stacked models.



(a) Binary



(b) Combined



(c) All

Figure 7.8: The confusion matrices of the stacked models for the different labelsets

7.1.4 The proposed decision tree

The proposed decision tree consists of three stages to predict unseen attempts. The first stage of the decision tree predicts if the attempt contains an error. If the attempt contains an error, the second stage predicts if the error is in the combined class of the combined labelset or not. If the attempt is predicted to be in the combined class, the third stage predicts the label of the attempt. This tree is also graphically visualized in figure 6.7.

To determine the models to use in the decision tree, hyperparameter tuning, balancing, and stacking was performed for the combined labelset, but without the zero errorclass to determine the model for stage two, and for the attempts labeled with labels ten to twenty, and sixty-six, which were combined in the combined labelset, to determine the model for stage three of the proposed decision tree. Although errorclass nine (no maixmal inhalation) was filed in the 'other: namely' field in the form, this is a criteria by Miller et al. (2005 [3] and thus treated as a single class when using the combined labelset. Refer to appendix C.3 for the results of the hyperparameter tuning, balancing and stacking.

Table 7.8 shows the models used in the three stages of the decision tree. The performance of the proposed decision tree is shown in table 7.9. Due to the set up of the decision tree, not all probabilities per predicted attempt could be calculated, making it impossible to calculate the precision at 100% recall. Therefore, the F1-score is used to compare the trade-off between precision and recall.

Labelset	Featureset	Model	Balancing technique	Precision	Recall	F1-score	Precision at 100% recall
Binary (stage 1)	Filtered features	SVM	ROS	0.857	0.864	0.861	0.678
Combined, without the zero errorclass (stage 2)	Filtered features	Boosted decision trees	ROS	0.343	0.338	0.332	0.127
Attempts with labels between 10 and 20, and 66 (stage 3)	Filtered features	Boosted decision trees	ROS	0.350	0.300	0.305	0.125

Table 7.8: The models used in the three stages of the proposed decision tree

	Precision	Recall	F1-score
Proposed decision tree	0.147	0.144	0.142

Table 7.9: The performance of the proposed decision tree

The proposed decision tree performed worse compared to the best single model using labelset 'all', with a difference in recall and F1-score of respectively 0.178 and 0.179.

The confusion matrix is presented in figure 7.9. The errorclasses are mostly confused with errorclass six (no maximal effort). Also, this errorclass is confused with the most errorclasses. When comparing to the confusion matrix of the best model trained and evaluated on labelset all (figure *c* of 7.7), we see that only four attempts of errorclass sixty-six (multiple errors) were classified right by the proposed decision tree, compared to all being classified right when using the best model for labelset 'all'. However, more attempts of the zero errorclass (technically correct attempts) were classified right using the proposed decision tree. Also, errorclass six (no maximal effort) was more confused with the other errorclasses, and the other errorclasses with errorclass six, when using the proposed decision tree.

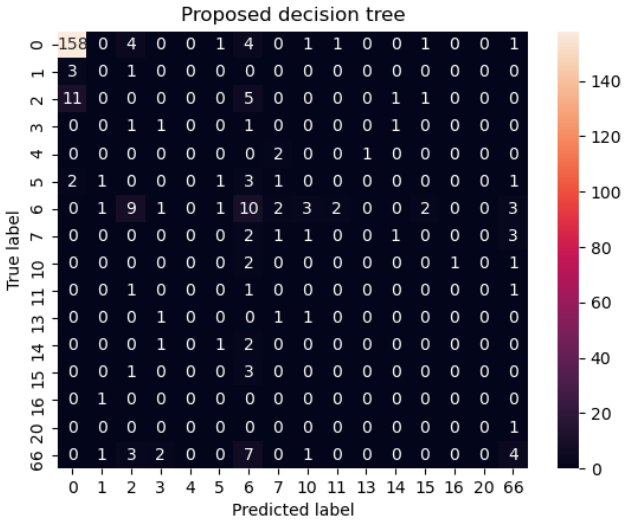


Figure 7.9: Confusion matrix of the proposed decision tree.

7.1.5 The best fit

The models which classify unseen spirometry attempts best for the three labelsets are presented in table 7.10.

Labelset	Featureset	Model	Balancing technique	Precision	Recall	F1-score	Precision at 100% recall
Binary	Filtered features	SVM	ROS	0.857	0.864	0.861	0.678
Combined	Spirometry parameters	SVM	ROS	0.584	0.525	0.547	0.134
All	Spirometry parameters	SVM	ROS	0.340	0.322	0.321	0.063

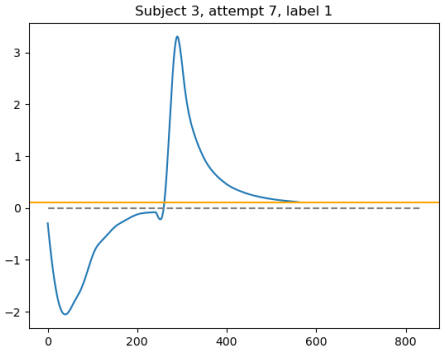
Table 7.10: The best performing models for the different labelsets, together with the featureset and balancing technique they perform best with.

The best results are achieved when using the SVM and balancing technique ROS. The featureset they performed best with is the filtered featureset (refer to appendix B.2 for the features) for the binary labelset, and the spirometry parameters for labelsets 'combined' and 'all'.

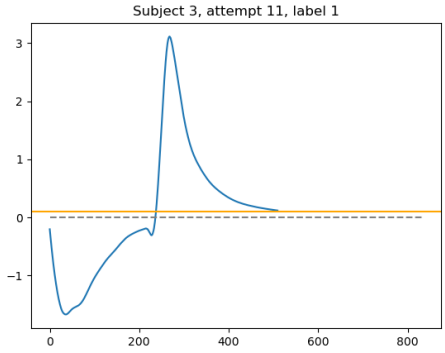
The time needed to classify unseen attempts, on a laptop with an intel core i5, and an 250 GB SSD, by the three models is negligible, excepting to form no problem in the real life system.

The confusion matrices of these best models are shown in figure 7.7 and discussed in section 7.1.3. It was seen that errorclass two (obstructed mouthpiece) was confused the most with the other errorclasses.

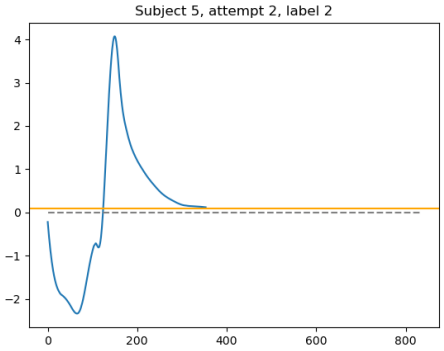
Figure 7.10 visualises the attempts which are classified wrong by the three best performing models.



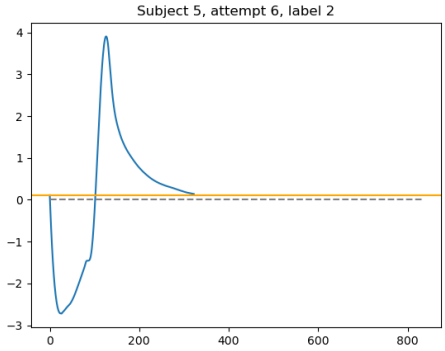
(a)



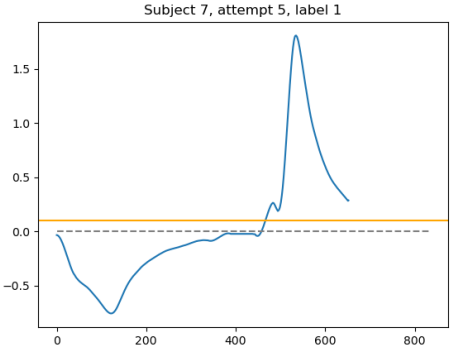
(b)



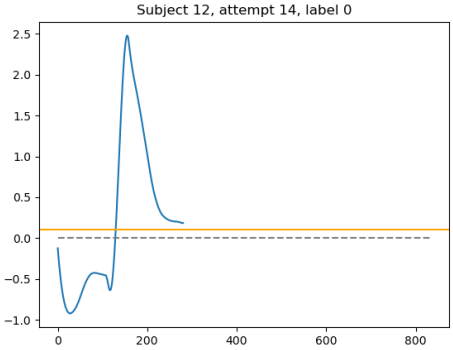
(c)



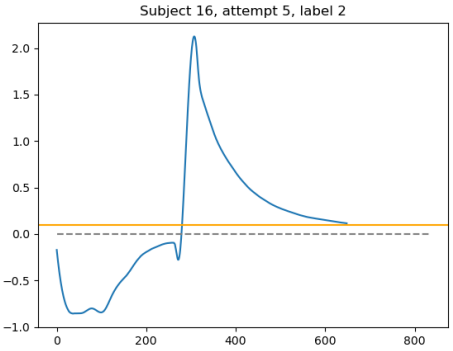
(d)



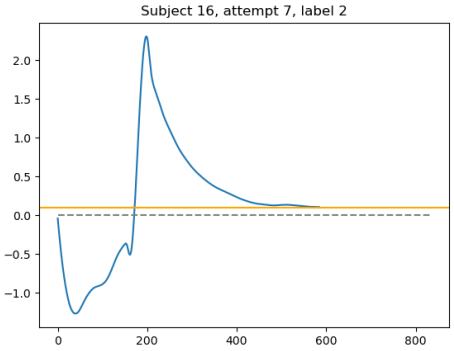
(e)



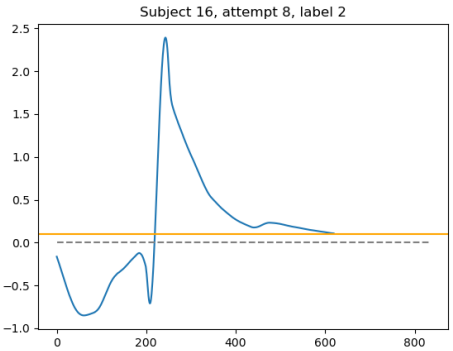
(f)


Subject 16, attempt 5, label 2

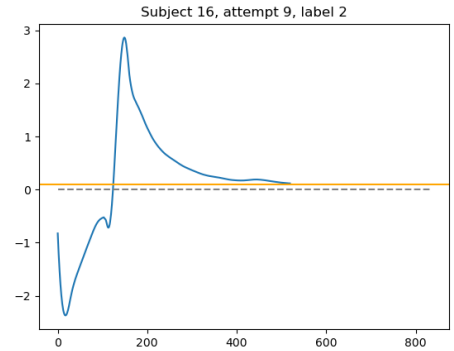
(g)


Subject 16, attempt 7, label 2

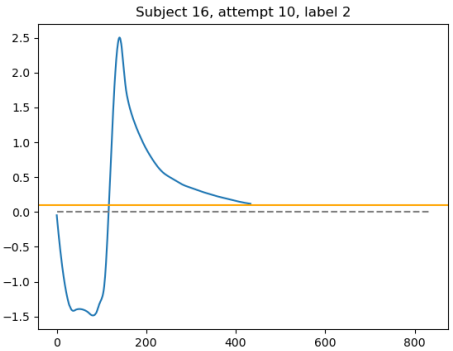
(h)


Subject 16, attempt 8, label 2

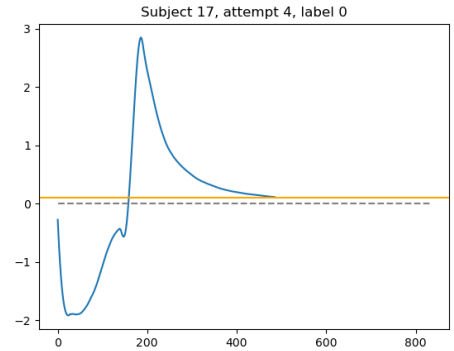
(i)


Subject 16, attempt 9, label 2

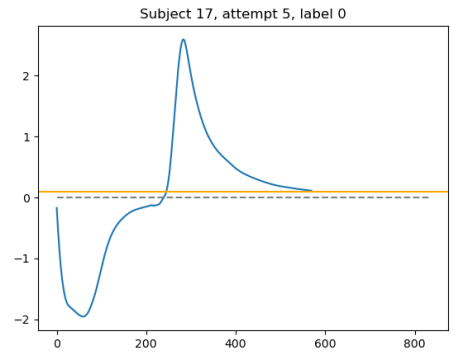
(j)


Subject 16, attempt 10, label 2

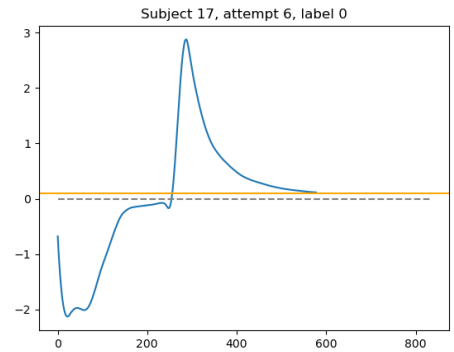
(k)


Subject 17, attempt 4, label 0

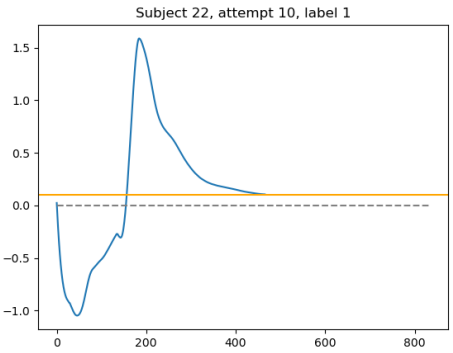
(l)


Subject 17, attempt 5, label 0

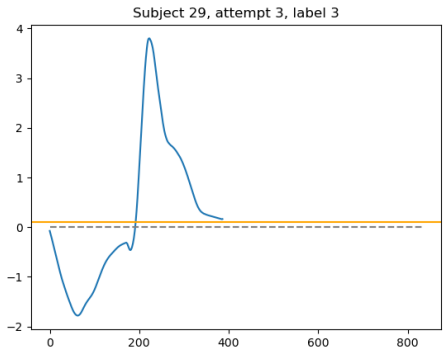
(m)


Subject 17, attempt 6, label 0

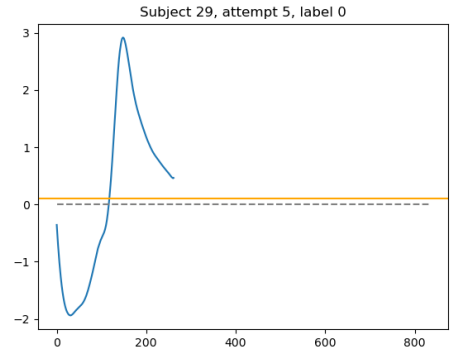
(n)



(o)



(p)



(q)

Figure 7.10: The misclassified attempts for all labelsets. The x-axis represents the time in 10^1 ms, while the y-axis represents the flow in liters per second. The dashed line represents the y-axis being zero, the orange line shows the y-axis being 0.1. The end of the exhalation is where the curve crosses this line.

The attempts are technically correct attempts (label zero), attempts with an unsatisfactory start (label one), attempts during which the mouthpiece was obstructed (label two), and attempts during which an extra breath was taken (label three). When looking at the attempts, one can see that all attempts have a dip in the inhalation before the exhalation, although small in cases *m* and *q*.

7.1.6 Including the data of the inter-annotation study

To evaluate if including more data increases the performance of the best performing models, the data of the inter-annotation study was included in the trainingset. The labels of the rater which training and experience level is closest to the professional labeling the data of the hospital experiments were used. The total number of attempts in the combined dataset is 500. After removing the outliers, the dataset

consists of 472 attempts. The label distribution of this combined dataset is shown in figure 7.11.

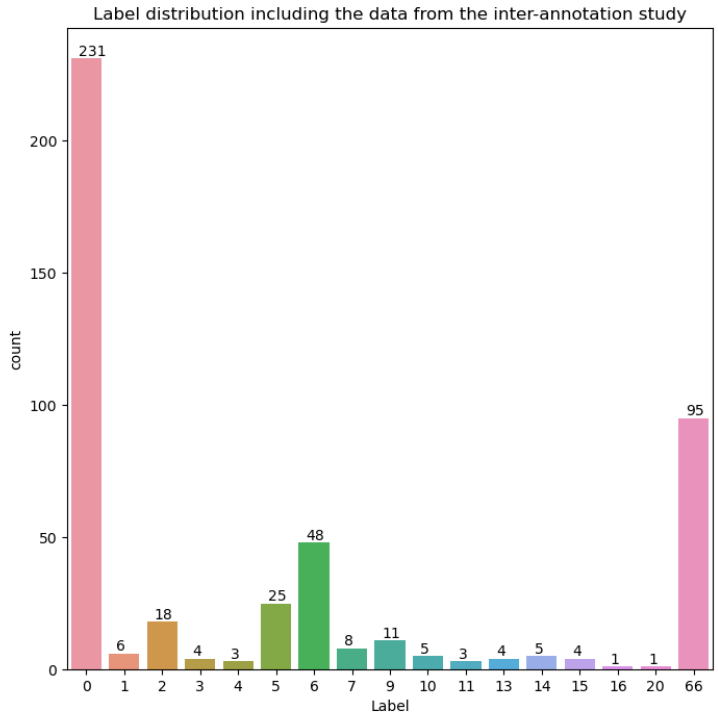


Figure 7.11: The label distribution when combining the dataset of the hospital experiments and the dataset of the inter-annotation study.

The participants of the inter-annotation study were students between eight-teen and thirty-four years old who were not trained in performing spirometry attempts. Also, they were coached by a non-professional in supervising spirometry attempts. The students made a lot of mistakes, and many attempts contained multiple errrors according to the rater.

The best performing models trained and evaluated on the data of the hospital experiments (refer to section 7.1.5) were used to compare the performance when including the data of the inter-annotation study, and when training and evaluating on the data of the inter-annotation study only.

The performance of the models trained on the data of the inter-annotation study only are presented in table 7.11 and show that the performance is in general lower than the performance when trained and evaluated on the data of the hospital experiments. However, the precision at 100% recall is higher for the labelsets 'combined' and 'all'.

Labelset	Featureset	Model	Balancing technique	Precision	Recall	F1-score	Precision at 100% recall
Binary	Filtered features	SVM	ROS	0.718	0.642	0.678	0.5
Combined	Spirometry parameters	SVM	ROS	0.354	0.350	0.341	0.167
All	Spirometry parameters	SVM	ROS	0.311	0.276	0.279	0.111

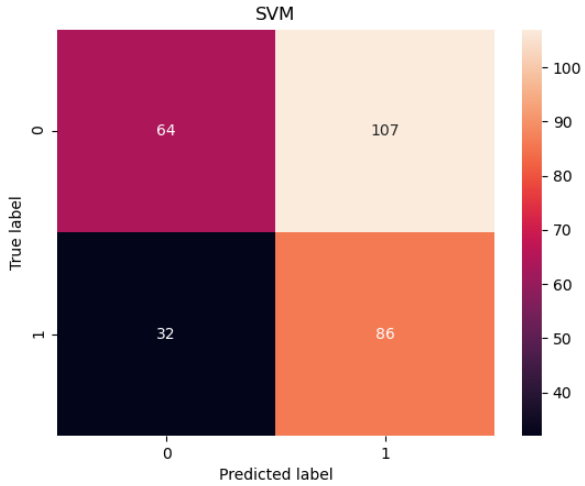
Table 7.11: The performance of the models trained and evaluated on the data of the inter-annotation study only. The models used are the best performing models trained and evaluated on the data of the hospital experiments.

When we look at the performance of the models trained on the data of the hospital experiments and the data of the inter-annotation study (table 7.12), we see that the performance of the models is lower than when trained on only the data of the inter-annotation study for the labelsets 'combined' and 'all'. However, for the binary labelset, the recall and precision at 100% recall are higher when using the data of both studies.

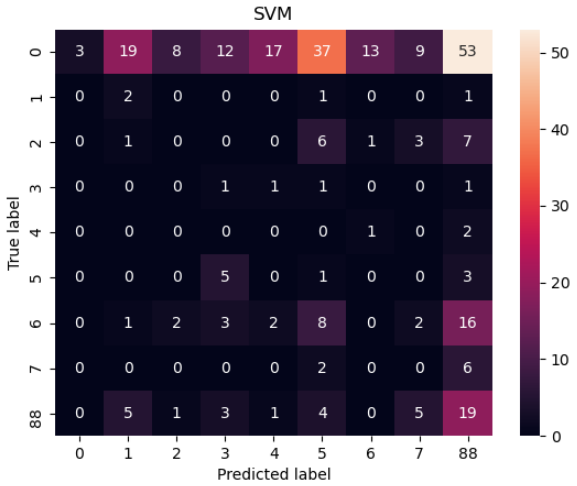
Labelset	Featureset	Model	Balancing technique	Precision	Recall	F1-score	Precision at 100% recall
Binary	Filtered features	SVM	ROS	0.446	0.729	0.553	0.541
Combined	Spirometry parameters	SVM	ROS	0.062	0.195	0.093	0.113
All	Spirometry parameters	SVM	ROS	0.023	0.102	0.037	0.063

Table 7.12: The performance of the models when including the data of the inter-annoation study in the trainingset.The models used are the best performing models when only training on the dataset of the short term study.

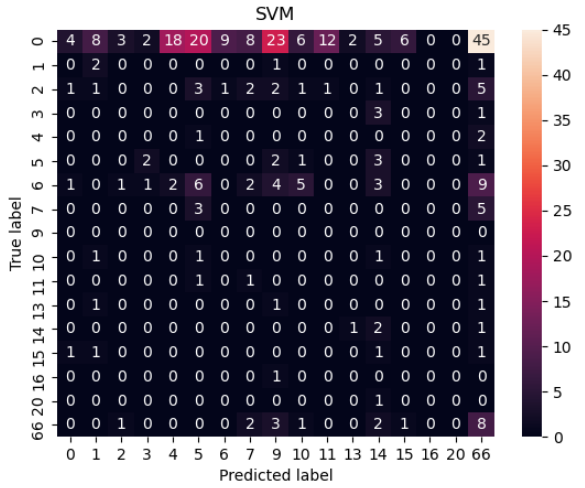
The confusion matrices of the models trained on the data of both studies are shown in figure 7.12.



(a) Binary



(b) Combined



(c) All

Figure 7.12: The confusion matrices of the best performing models for the different labelsets with the data of the inter-annotation study added to the trainingset.

These matrices show that a lot of attempts from the errorclasses are confused with the zero errorclass (technically correct attempts). Also, the attempts from the combined errorclass (eighty-eight) when using the combined labelset, and the attempts with multiple errors (label sixty-six) when using labelset 'all', were confused with the other classes, while this was less the case when the models were only trained on the data of the hospital experiments. Additionally, errorclass six (no maximal effort) which was classified well when only training on the hospital experiments, is now not classified right at all when using the labelsets 'combined' and 'all'. These differences are also seen in the confusion matrices when trained and evaluated on

the data of the inter-annotation study only (figure C.1 in the appendix).

When comparing the attempts that were misclassified by the three models when only trained on the data of the hospital experiments (figure 7.10) and when trained on both the data of the hospital experiments and the inter-annotation study, it was found that only five out of seventeen overlapped. These were the attempts shown in figures *a*, *b*, *c*, *d*, and *q* of figure 7.10. No reason for only these five being misclassified could be deduced from the curves.

7.1.7 Comparison to a rule-based approach

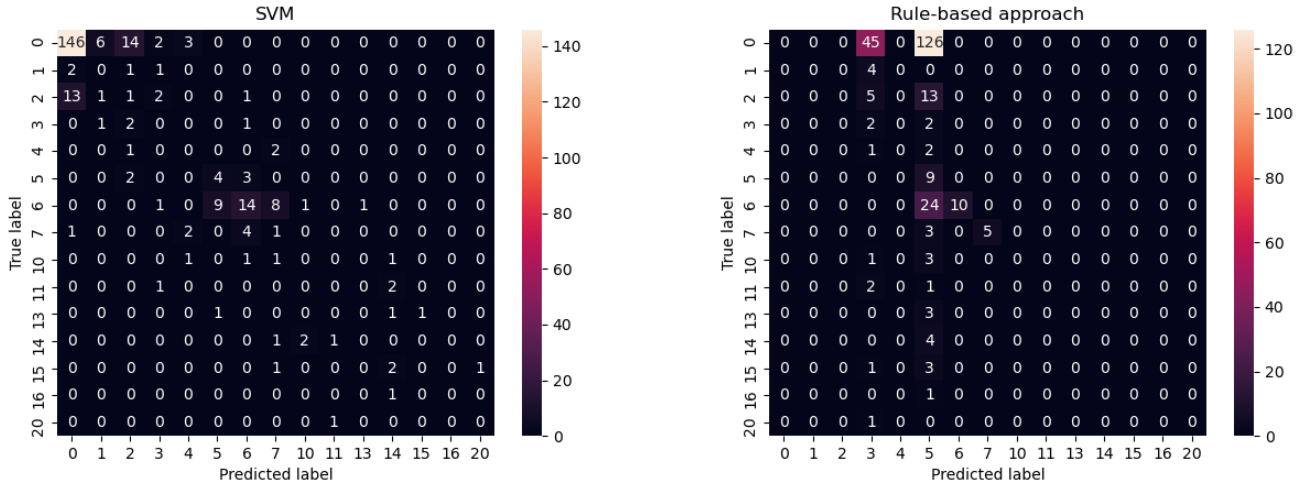
The best performing model trained and evaluated using labelset 'all' is compared to the rule based approach explained in section 6.1.4. The attempts with multiple errors were excluded from the comparison, as otherwise all attempts not meeting rules one, two, four, and five would have been classified by the rule-based approach as containing multiple errors, resulting in an unfair comparison.

The results of both approaches are presented in table 7.13. The rule-based approach outperforms the machine learning approach by a difference in precision, recall, and F1-score of respectively 0.187, 0.06, and 0.014.

Approach	**Precision**	**Recall**	**F1-score**
Rule-based approach	0.425	0.26	0.226
Machine learning approach	0.238	0.2	0.212

Table 7.13: The performance of the rule-based and machine learning approach. The attempts with multiple errors are neglected in training and evaluation.

The confusion matrices of both approaches are presented in figure 7.13. While the machine learning approach classifies the zero errorclass (technically correct attempts) well, the rule-based approach misclassifies all the attempts from this errorclass. Additionally, only twenty-six out of 271 attempts were classified right by the rule-based algorithm, compared to 166 by the machine learning approach. As the rule-based approach is only able to detect errors one (unsatisfactory start), three (extra breath taken during the attempt), five (terminated too early), six (no maximal effort), seven (cough), and eight (glottis closure), the attempts of the rest of the errorclasses are misclassified. Most confusion was with errorclass five (terminated too early). This is due to the set up of the comparison; when none of the labels given by the rule-based approach overlapped with the real label, the most important error detected by the rule-based approach was assigned to this attempt, which was 185 out of 245 times none of the labels overlapped.



(a) The machine learning approach

(b) The rule-based approach

Figure 7.13: The confusion matrices of the two approaches. The attempts with multiple errors were neglected in training and evaluation.

7.2 Inter-annotation study

To determine the agreement in labeling by multiple professionals, an inter-annotation study was performed. Refer to section 6.2 for details.

7.2.1 Data gathering

Thirteen adults participated in this study, which took one week. Due to the COVID-19 crisis, all professional locations were not accessible, which resulted in performing the experiment in a student room. To make the location a public space as much as possible, the curtains were open, and the door unlocked. Furthermore, the participants were asked if they agreed with this setup. None of the thirteen participants disagreed.

During one test, the cameras did not work. These attempts were neglected in further processing. Besides, the camera recording the side view of another participant did not work during two attempts. As the other camera recorded the attempt, these attempts were included in the dataset. Besides these two problems, the study went as planned.

Three professionals labeled the data of thirteen adults, which summed up to 191 attempts in total. Professional one is a medical student, trained during an internship and courses, the other two professionals work at the MST hospital in Enschede as technical physicians. They have some years of experience in supervising and assessing spirometry tests.

7.2.2 Determination of the agreement

The Cohen's kappa score was calculated for the inter- and intra-agreement for all combinations of rounds. This was done for the three labelsets used in designing the error detection algorithm.

The outliers were removed. Additionally, one of the professionals told in hindsight that he did not always assigned label five (terminated too early), as he would assign this label to almost all of the attempts. As the other professionals did not know this before the first round of labeling, this errorclass was neglected in the calculation of the agreements.

Tables 7.14, 7.15, and 7.16 show the inter- and intra-rater agreements for the three labelsets. The intra-rater agreements are in bold for distinctiveness.

The inter-rater agreements range from -0.123 to 0.380 for the binary labelset, -0.035 to 0.301 for the combined labelset, and -0.035 to 0.304 for labelset 'all'. These agreements are all negative, none or minimal agreements (refer to table 6.5 for the interpretation). In general, the agreement decreases when the amount of classes the agreement is based on increases. However, the agreements between raters two and three are higher when using labelset 'all' compared to using labelset 'combined'.

	Rater 1 Round 1	**Rater 1 Round 2**	**Rater 2 Round 1**	**Rater 2 Round 2**	**Rater 3 Round 1**
Rater 1 Round 2	**0.865**				
Rater 2 Round 1	0.066	0.080			
Rater 2 Round 2	0.135	0.153	**0.780**		
Rater 3 Round 1	-0.078	-0.054	0.380	0.311	
Rater 3 Round 2	-0.123	-0.094	0.275	0.198	**0.860**

Table 7.14: The kappa score representing the inter- and intra-agreement, using the binary labelset. The kappa scores of the intra-agreement are in bold for distinctiveness.

	Rater 1 Round 1	**Rater 1 Round 2**	**Rater 2 Round 1**	**Rater 2 Round 2**	**Rater 3 Round 1**
Rater 1 Round 2	**0.862**				
Rater 2 Round 1	0.001	0.017			
Rater 2 Round 2	0.047	0.065	**0.724**		
Rater 3 Round 1	-0.022	-0.015	0.301	0.217	
Rater 3 Round 2	-0.035	-0.029	0.156	0.127	**0.648**

Table 7.15: The kappa score representing the inter- and intra-agreement, using the combined labelset. The kappa scores of the intra-agreement are in bold for distinctiveness.

	Rater 1 Round 1	**Rater 1 Round 2**	**Rater 2 Round 1**	**Rater 2 Round 2**	**Rater 3 Round 1**
Rater 1 Round 2	**0.862**				
Rater 2 Round 1	0.001	0.009			
Rater 2 Round 2	0.047	0.057	**0.725**		
Rater 3 Round 1	-0.022	-0.015	0.304	0.241	
Rater 3 Round 2	-0.035	-0.028	0.165	0.153	**0.660**

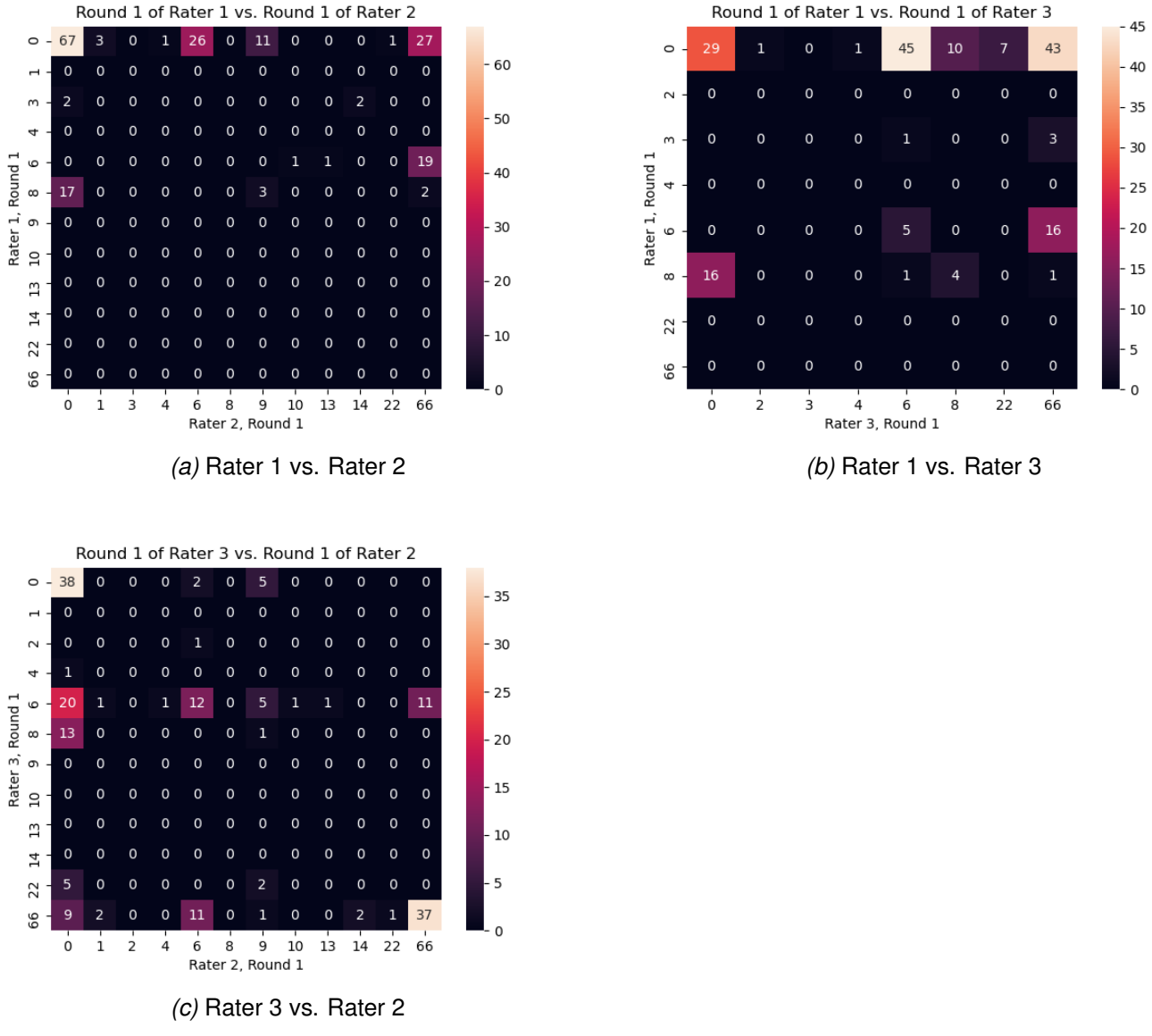
Table 7.16: The kappa score representing the inter- and intra-agreement, using labelset 'all'. The kappa scores of the intra-agreement are in bold for distinctiveness.

The inter-rater agreements between rater two and three are the highest, and the agreements between rater one and three the lowest, independent of which labelset was used. The inter-rater agreements between rater one and three are below zero, meaning the agreement is less than it would be by chance.

The intra-rater agreements range from 0.780 to 0.865 for the binary labelset, from 0.648 to 0.862 for the combined labelset, and from 0.660 to 0.862 for labelset

'all'. These agreements can be interpreted as moderate to strong according to table 6.5. The intra-rater agreement of rater one is the highest, the lowest intra-rater agreement is by rater two for the binary labelset, and by rater three for the other two labelsets.

The confusion matrices comparing the first rounds of the raters are shown in figure 7.14. All confusion matrices can be found in Appendix E.



(a) Rater 1 vs. Rater 2

(b) Rater 1 vs. Rater 3

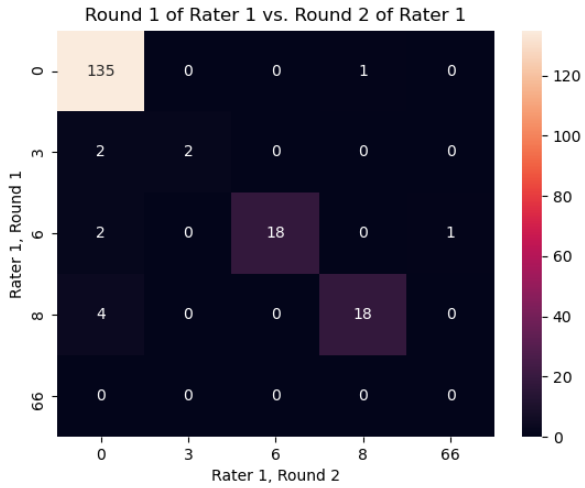
(c) Rater 3 vs. Rater 2

Figure 7.14: The confusion matrices comparing the labels given by the three raters during the first round, using labelset 'all'.

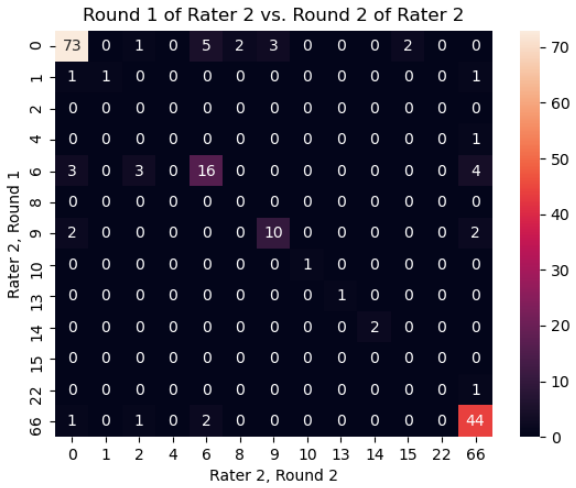
Confusion matrices *a* and *b* show that the confusion between the labels of rater one and the labels of the other two raters is primarily with the zero errorclass (technically correct attempts). Additionally, rater one labeled none of the attempts as

containing multiple errors (label sixty-six), while the other two raters did. Besides, raters one and three rated respectively twenty-two and fourteen attempts as label eight only (wrong position), while rater two labeled none of the attempts as containing this error. When comparing the first rounds of raters two and three (confusion matrix *c*), we observe the most confusion between attempts labeled with label six (no maximal effort) by rater three, which were labeled mainly as zero (technically correct attempt), nine (no maximal inhalation), or sixty-six (multiple errors) by rater two. Additionally, attempts during which the person did not have a right position (label eight) according to rater three, were mainly labeled by rater two as a technically correct attempt (label zero). Also, attempts labeled as technically correct (label zero) by rater two, were mainly labeled by rater three with labels six (no maximal effort), label eight (wrong position), or were assigned multiple labels (label sixty-six). Finally, the attempts labeled with multiple labels (label sixty-six) were mainly confused with labels six (no maximal effort) and zero (technically correct attempt).

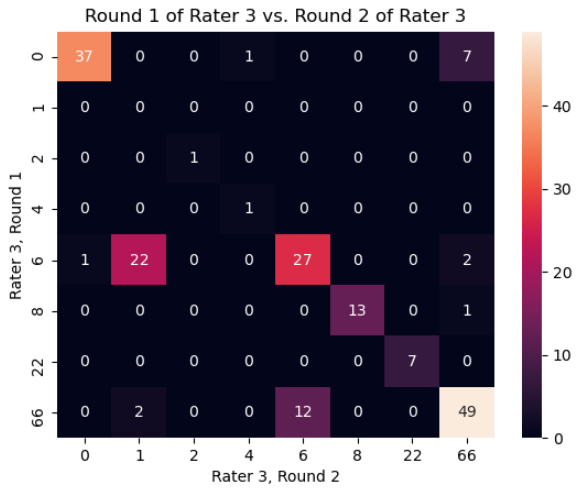
The confusion matrices comparing the labels given by the three raters over time are shown in figure 7.15. The confusion between the two rounds of rater one (confusion matrix *a*) is mostly in attempts that were labeled as technically correct (label zero) in round two, but were labeled as containing an error in round one. The main confusion between the two rounds of rater two were in attempts which were labeled as technically correct (label zero) during round one, but labeled as containing an error during round two, and attempts which were labeled as containing multiple errors in round two, but were labeled with a single label in round one. For rater three, the main confusion is between errorclasses one (unsatisfactory start), and six (no maximal effort), and errorclass sixty-six (multiple errors), zero (technically correct attempts), and six (no maximal effort).



(a) Round 1 vs. Round 2 of Rater 1



(b) Round 1 vs. Round 2 of Rater 2



(c) Round 1 vs. Round 2 of Rater 3

Figure 7.15: The confusion matrices comparing the labels given by the two raters during the first and second round, using labelset 'all'.

7.3 Comparison of coaching by a metaphor versus by a professional

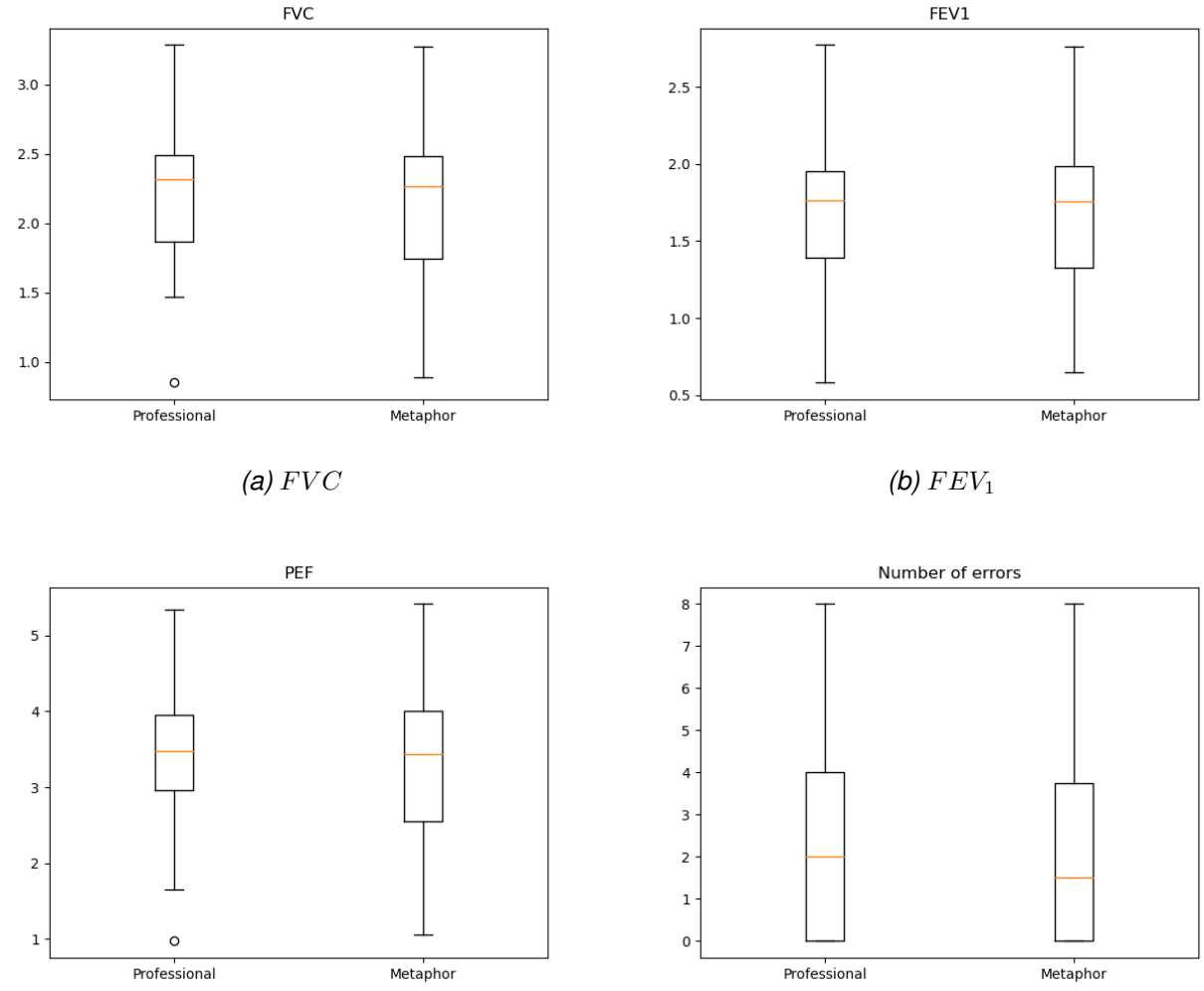
To assess if there is a significant difference in quality of the attempts coached by a professional or by a metaphor, the FVC, FEV_1, PEF values, and the number of errors occured during the hospital experiments, explained in section 6.1.1, were compared.

7.3.1 Preprocessing

Before the values were compared, the data was preprocessed. As for the comparison of the FVC, FEV_1, and PEF values only the attempts without an error are used, the attempts which were not technically correct were removed from the dataset. If a subject did not blow technically correct attempts when either coached by the professional or the metaphor, the attempts from this subject are excluded from the comparison. This resulted in attempts from twenty-six subjects to compare.

7.3.2 Data exploration

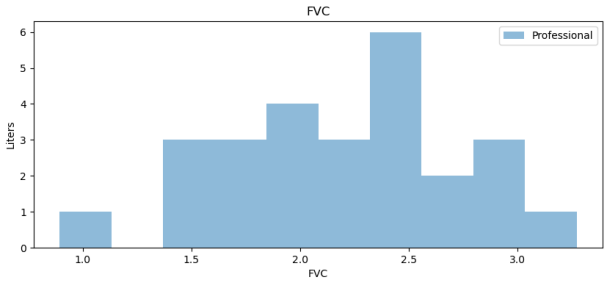
To visualize the data distributions, boxplots and histograms were created and presented in figures 7.17 and 7.16. These visualizations show that the distributions of the two ways of coaching are alike each other. The mean values differ little, except for the mean number of errors, where the mean value when coached by a metaphor is lower. Furthermore, the variability in FVC and PEF is less when the children were coached by a professional compared to by a metaphor, while the other variabilities are approximately similar.



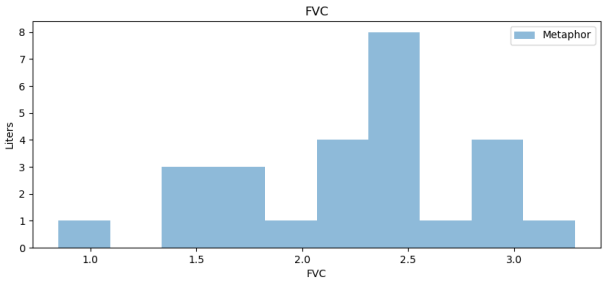
(a) FVC *(b)* FEV_1

(c) PEF *(d)* Number of errors

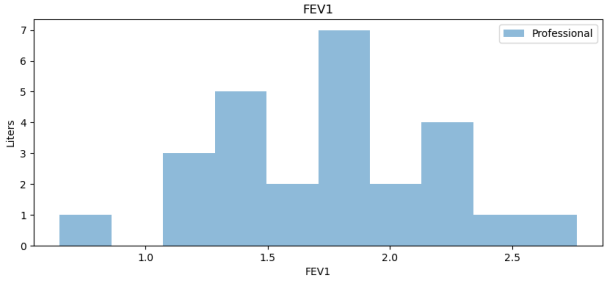
Figure 7.16: Boxplots showing the distributions of the attempts coached by a professional and by a metaphor.



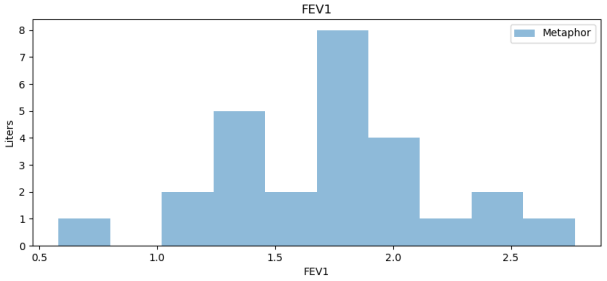
(a) FVC, when coached by a professional



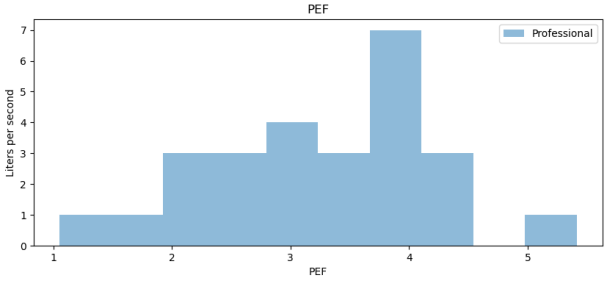
(b) FVC, when coached by a metaphor



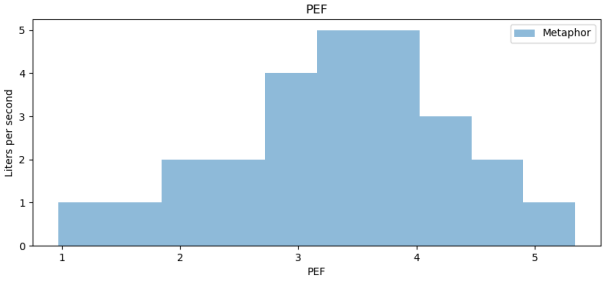
(c) FEV_1, when coached by a professional



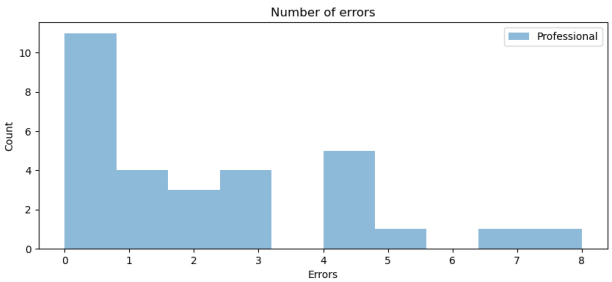
(d) FEV_1, when coached by a metaphor



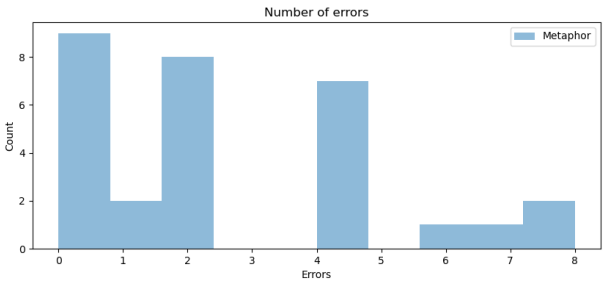
(e) PEF, when coached by a professional



(f) PEF, when coached by a metaphor



(g) Number of errors, when coached by a professional



(h) Number of errors, when coached by a metaphor

Figure 7.17: Histograms showing the distributions of the attempts coached by a professional and by a metaphor.

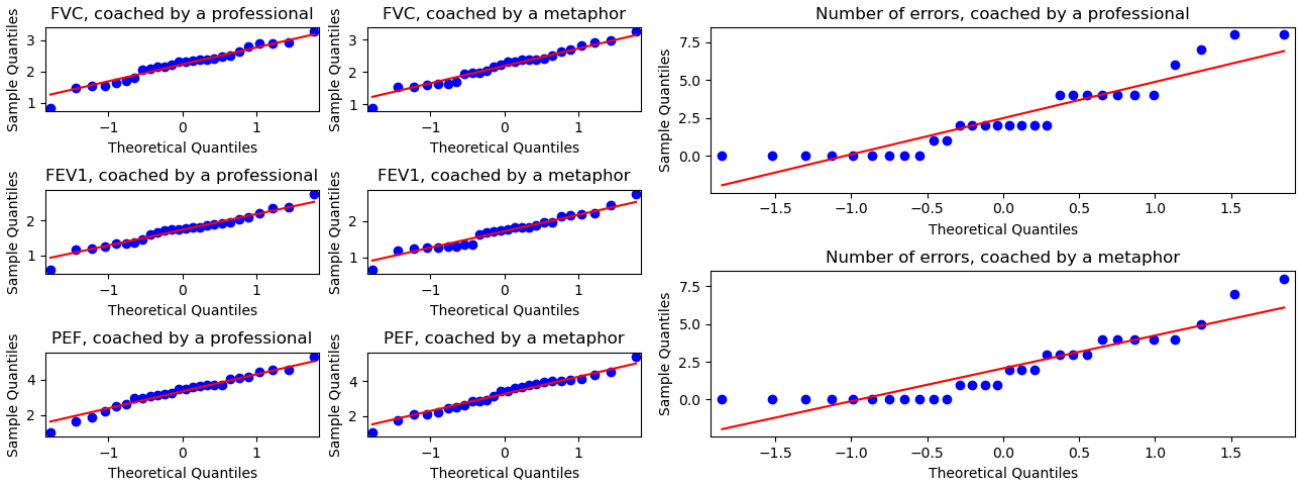
7.3.3 Data analysis

First, a Shapiro-Wilk test was performed and QQ-plots were created to determine if the data is normally distributed. The output from the Shapiro-Wilk test can be

found in table 7.17, the QQ-plots in figure 7.18. The threshold for being normally distributed was set at an alpha of 0.05.

Coaching manner	**Metaphor**		**Professional**	
Statistics	***D(26)***	***p***	***D(26)***	***p***
FVC	0.980	0.877	0.973	0.705
FEV_1	0.973	0.703	0.979	0.860
PEF	0.982	0.920	0.981	0.895
Statistics	***D(30)***	***p***	***D(30)***	***p***
Number of errors	0.852	0.001	0.861	0.001

Table 7.17: The outcome of the Shapiro-Wilk tests, showing the spirometry parameters are normally distributed, while the number of errors is not.



(a) QQ-plots of the spirometry parameters

(b) QQ-plot of the number of errors

Figure 7.18: QQ-plots showing that the spirometry parameters are normally distributed, while the number of errors is not.

The FVC, FEV_1, and PEF values did not deviate significantly from normal according to the Shapiro-Wilk test; respectively $D(26)$=0.980, p = 0.877, $D(26)$=0.973, p = 0.703, $D(26)$=0.982, p = 0.920 for the metaphor values, and $D(26)$=0.973, p = 0.705, $D(26)$=0.979, p = 0.860, $D(26)$=0.981, p = 0.895 for the values blown when coached by the professional. However, when evaluating the number of errors, we see that number of errors when coached by a metaphor, $D(30)$=0.852, p=0.001, and when coached by a professional, $D(30)$, p=0.001, are both significantly not normal. These results are strengthened by the QQ-plots, showing that the FVC, FEV_1, and

PEF values are normally distributed, as the red line representing the expected line for a normal distribution is followed. The QQ-plot of the number of errors shows a big deviation from this line, presenting that this distribution is not normal.

As the number of errors are not normally distributed, the Wilcoxon test was used to compare the two coaching manners. For the comparison of the FVC, FEV_1,and PEF values, a paired sample t-test was conducted. The output of these tests can be found in table 7.18.

Statistics (paired sample t-test)	**Mean**		**Standard deviation**		**95% CI**	**t(50)**	**p**
Coaching manner	***Met.***	***Prof.***	***Met.***	***Prof.***			
FVC	2.188	2.224	0.537	0.536	[0.007, 0.919]	1.327	0.197
$FEV1$	1.725	1.736	0.452	0.447	[0.097, 0.994]	0.518	0.609
PEF	3.270	3.336	0.970	0.963	[0.004, 0.953]	1.164	0.255
Statistics (Wilcoxon test)	**Mean**		**Standard deviation**		**95% CI**	**t(61)**	**p**
Coaching manner	***Met.***	***Prof.***	***Met.***	***Prof.***			
Number of Errors	2.067	2.5	2.175	1.391	[0.002, 0.947]	61.5	0.169

Table 7.18: The outcome of the paired sample t-test and Wilcoxon test. 'Met.' stands for 'metaphor', and 'Prof.' stands for 'professional'.

On average, the FVC value was found to be higher when coached by a professional (M = 2.224, SE=, 0.536) than when coached by a metaphor (M = 2.188, SE = 0.537). This difference, 0.001, BCa 95% CI [0.007, 0.919], was not significant $t(50)$ = 1.327, p = 0.197. The FEV_1 value was found to be higher, on average, when coached by a professional (M = 1.736, SE=, 0.447) than when coached by a metaphor (M = 1.725, SE = 0.452). This difference, 0.011, BCa 95% CI [0.097, 0.994], was not significant $t(50)$ = 0.518, p = 0.609. When evaluating the PEF value, it was found to be higher, on average, when coached by a professional (M = 3.336, SE=, 0.963) than when coached by a metaphor (M = 3.270, SE = 0.970). This difference, 0.066, BCa 95% CI [0.004, 0.953], was not significant $t(50)$ = 1.164, p = 0.255. The means of the number of errors for the two coaching manners was evaluated using the Wilcoxon test, as the distribution was found to be not normal. On average, the number of errors were found to be higher when coached by a professional (M = 2.5, SE= 1.391) than when coached by a metaphor (M = 2.067, SE = 2.175). This difference, 0.433, BCa 95% CI [0.002, 0.947], was not significant $t(61)$ = 61.5, p = 0.169.

Chapter 8

Discussion and recommendations

8.1 Error detection algorithm

One of the goals of this research was to design and evaluate an error detection algorithm for home spirometry measurements, which is able to automatically detect errors in spirometry attempts. Normally, a professional supervising spirometry attempts at the hospital would assess these attempts. However, no professional is available at home. Therefore, the error detection algorithm should support the quality of the home spirometry attempts. When the errors can be detected well, the quality of the home spirometry attempts increases, making it possible to monitor the asthma patients at home. Besides, feedback can be provided to the patient, based on the attempt, which makes it easier for the patient to blow acceptable attempts.

In this research different classifiers, featuresets, labelsets, and machine learning techniques are used to determine the best algorithm to detect errors in spirometry attempts. This section discusses the results.

8.1.1 Dataset

The dataset used to train and evaluate the different models consists of the data gathered during the hospital experiments, explained in section 6.1.1. During this study, thirty children performed two tests; one in which they were coached by a professional, and one in which they were coached by metaphors, which are described in section 2.4. Both tests were included due to the small dataset, although only using the attempts coached by a metaphor would mimick the real life situation better. When more data is available, only these attempts should be included.

8.1.2 Outlier removal

To reduce the noise in the dataset as complete as possible, outliers based on visual inspection, and based on the FEV_1 and FVC values, were removed. The outliers were from ten different classes, and thus were not attempts containing a specific error. Although these outliers are not used as trainingdata for the algorithm, a system could be designed in which the outliers are detected by the heuristics used in this research to detect the outliers before using the algorithm, making it possible to provide feedback after these attempts. In this way, the noise in training the algorithm is reduced, while still being able to give general feedback to all attempts.

Another option would be to see the outliers as an error. For example, a low FVC value may point to the error that no maximal effort was shown. Not removing the outliers which represents errors could improve the performance of the algorithm. Therefore, a recommendation for future research is to decide in consultation with a professional if an attempt is an outlier, or useful in training the algorithm.

8.1.3 Featuresets

Five different combinations of featuresets are used in this research as input for the machine learning models. These featuresets consisted of timeseries features, spirometry parameters, and the age and sex of the children.

The performance of the models was best when the spirometry parameters and the age and sex of the children were used as input for labelsets 'combined' and 'all'. This is remarkable, as the expectation was that the time-series features would contain more information about what errors are in the attempts, as these features include for example the number of peaks in the data; two peaks may reveal a cough during the attempt. However, as only the flow-time curve was used to extract time-series features from, it could be that not enough, or not the right, information was extracted. Therefore, it is recommended to also extract features from the flow-volume curve and the volume-time curve, instead of only from the flow-time curve, as this will add information. Additionally, smoothing the data reduced noise, however it could have reduced information as well. Therefore, training on the unsmoothed data is recommended to evaluate.

The best performing model for the binary labelset used the filtered features as input, showing that the time-series features extracted contained valuable information about if an attempt contains an error or not. The expectation is that extracting features from the named curves will improve the performance of the model.

The age and sex was used to help the model distinguish between attempts which for example have a low FVC due to an error, or due to the young age of the child.

The age of the children was only known in years. More information would be provided by representing the age in for example days, as this is more accurate.

The height of the children was unknown, however this is of influence on the lung capacity of the children. The predicted PEF and FVC of the children are known, as these are needed to calibrate the metaphors. These values are based on the age, height, and sex of the children and will add valuable information if these are used as features instead of, or as addition to, the age and sex of the children. The values and predicted values can also be combined into one feature by using the percent predicted value instead.

8.1.4 Labelsets

Three labelsets are used in this research. A binary labelset with a non-errorclass and an errorclass, a combined labelset in which the errors from all 'other, namely:' classes are combined, and labelset 'all' in which all errorclasses were seen as single classes.

The best performing model using the combined labelset achieved a recall of 0.525 and a precision at 100% recall of 0.134. The expectation is that the performance of the combined labelset could be improved by combining errorclasses based on the part of the manoeuvre the error occured in, or by combining the errors for which the same feedback shoud be given, as these errors are more alike. If this grouping of classes performs well, the system will be able to provide more specific feedback, compared to the grouping of errorclasses performed in this research.

8.1.5 Hyper-parameter tuning and balancing

The models were trained using different hyperparameters to find the best hyperparameterset, and two balancing technique to produce balanced datasets. The first balancing technique was SMOTE, which adds information by creating a new sample along the line between two existing points. The second balancing techinque was ROS, which simply duplicates random datapoints.

Balancing the dataset improved the performance of the models slightly, with a bigger mean improvement in recall compared to the mean improvement in the precision at 100% recall. It is remarkable that the performance did not improve substantially as the data was very skewed before balancing the dataset. From the confusion matrices of the best models before and after balancing, we deduced that before balancing, the most confusion was with errorclass zero (technically correct attempts), errorclass six (no maximal effort), and errorclass eighty-eight (combined errorclass) when using the combined labelset. These classes were the bigger classes. After

balancing the dataset, the confusion was more spread over classes. However, as there is still a lot of confusion, the performance did not improve to a large extent.

Balancing techinque ROS outperformed SMOTE. One reason could be that the information added by SMOTE for one errorclass overlapped with the information from other errorclasses, as this techinque does not take into account the other errorclasses when creating new samples. To improve the performance of SMOTE, this technique can be used in combinations with for example TomekLinks, which removes datapoints which have the smallest distance to datapoints from other classes.

8.1.6 Stacking the models

The output of the best models selected after hyperparameter tuning and balancing were combined and fed into a final estimator. The expectation was that this would increase the performance compared to the performance of the single models, as the information from the single models is combined. However, the performance decreased. This could be due to a high correlation between the models, as this would mean little information is added by combining the models. However, the correlation between the single models ranged from 0.563 to 0.583 for the binary labelset, from 0.234 to 0.434 for the combined labelset, and from 0.239 to 0.461 for labelset 'all', which is a negligible to moderate correlation [66]. Another reason could be a high performance on the trainingset, showing that the stacked models overfitted. The performance when evaluating the models on the trainingset (recall of 0.998, 0.852, 0.868 for respectively labelsets 'binary', 'combined', and 'all', and a precision at 100% recall of respectively 0.998, 0.624, and 0.626) show that the stacked models indeed overfitted. Due to the overfitting, the model was extremely good in classifying the attempts from the trainingset, but therefore performed less well on classifying unseen attempts. The overfitting on the trainingset should be reduced, by for example using a cross-validation method with a higher number of folds. Another option would be to use another final model. In this research, a decision tree is used. However, other classifiers may be better suitable for the data and result in a better performance of the stacked model. A final recommendation is to evaluate if other ensemble methods such as boosting, or bagging[1], improve the performance compared to the best performing single models in this study.

8.1.7 Proposed decision tree

The proposed decision tree consists of three stages. The first stage predicts if the attempt is flawless or not. If not, the second stage predicts if the error in the attempt

[1] https://blog.statsbot.co/ensemble-learning-d1dcd548e936

belongs to one of the single classes, or to the combined class, of the combined labelset. If this stage predicts it belongs to the combined class, the third stage predicts its final label.

To determine which model to use for the different stages, the models were optimized by hyper-parameter tuning, balancing, and stacking.

The proposed decision tree performed worse than the best model for labelset 'all', with a recall of 0.144 compared to 0.322. Although more of errroclass zero (technically correct attempts) were classified right by the decision tree, more attempts of the other classes were classified right by the single model, which are the more important classes to classify right. A reason for the low performance is that errors made in previous stages of the model can not be corrected by the following stages. Therefore, the confusion by the single stages accumulate, resulting in a low overall performance. The recommendation is to only use this model when the single stages perform perfectly.

8.1.8 The best fit

The best results met in this research are classifiers which are able to classify unseen attempts with a recall of 0.864, 0.525, and 0.322, for the labelsets 'binary', 'combined', and 'all', respectively. The performance represented by the precision at 100% recall of the best performing models was respectively 0.678, 0.134, and 0.063 for the labelsets 'binary', 'combined', and 'all'. However, the best models are selected based on the recall score. When we base this selection on the precision at 100% recall, the best performing model for labelset 'all' is different. This would be the RBFNN, trained on a imbalanced dataset of spirometry parameters (0.065).

According to a professional, the performance when using the binary labelset is accurate enough the be applied in a real life system. However, the performance of the moduls when using the labelsets 'combined' and 'all' is not good enough. An option is to use the binary prediction to determine if an attempt should be repeated, but base the feedback on the 'combined' or 'all' labelset.

One reason for the low performance when using the labelsets 'combined' and 'all' is that the correlation between attempts from different error classes is very high. This makes it hard for the classifiers to distinguish attempts from different classes. Besides, the dataset used was very small. The expectation is that extending the dataset will improve the performance as more data is added to the model, making it less severe for the model to distinguish between the errorclasses although the correlation is high.

When we compare our results to the results from Luo et al. (2017) [57] who evaluated a classifier predicting four common errors (early termination, cough, variable

flow, extra breath), we see that the algorithm proposed in the present study performed worse; they met an F1-score of 0.87 on average, while the F1-score of the best classifier to classify the errors using labelset 'all' is 0.321. However, the dataset in this research contains more errorclasses. Furthermore, the dataset used by Luo et al. (2017) [57] consisted of 1314 to 5728 curves per error class, while our data consisted of 309 attempts in total, and twelve out of the sixteen error classes were represented by less than ten attempts. The expectation is that when more data is included in the dataset, the performance of the algorithm proposed in this research will increase. Including more data will handle the high correlation between the attempts from different error classes in our dataset better and improve the performance.

The time needed to classify unseen data is a negligible amount of time, using a moderate laptop with an Intel core i5, and an 250 GB SSD. Therefore, this will form no problem in the real life system.

Misclassified attempts

The best performing models misclassified seventeen attempts, indepedently of which labelset was used. Refer to figure 7.10 for the curves of these attempts. The attempts were from errorclasses zero (technically correct attempts), one (unsatisfactory start), two (obstructed mouthpiece), and three (an extra breath taken during the attempt). This shows that the line between these errorclasses and the zero errorclass was thin in the data gathered during the hospital experiments. All attempts had a dip in the inhalation part. This part was covered by the filtered featureset, however only covered by the spirometry parameter Forced Inspirational Vital Capacity ($FIVC$) in the spirometry parameterset. The expectation is that including more features focusing on the inhalation part will improve the classification of these attempts.

It was expected that it would follow from the inter- and intra-rater agreements as well that these classes were difficult to distinguish. Although these classes are confused with other classes between professionals, (refer to figure 7.14), these are not exceptionally confused compared to other classes. Besides, when looking at the confusion matrices of the intra-rater agreement (refer to figure 7.15), errorclass one (unsatisfactory start) was not confused between the two round of rater two, but was confused by rater three, and not in the labelset of rater one. Errorclass two (obstructed mouthpiece) was not confused by rater three, but was confused by rater two, and again not in the labelset of rater one. No attempts were labeled as containing error three (an extra breath taken during the attempt) by rater two and three, and this class was not confused between the two rounds by rater one. These results show that it is not more difficult to distinguish these classes compared to the

other classes in the dataset by the professionals.

The expectation was that it would be difficult to classify attempts during which a flow leak occured as well, as this only leads to a small deviation in the data. Only three attempts with this error where available in the dataset. All attempts were indeed misclassified when using the labelsets 'combined' and 'all'. However, these were classified right when using the binary labelset. This shows that this errorclass is distinguishable from the zero errorclass representing flawless attempts, when using two classes, but not from errorclasses two (obstructed mouthpiece), and seven (cough), as these were the errorclasses these attempts were classified as.

When the data of the inter-annotation study was included in the dataset, only five attempts misclassified when training on only the data from the hospital experiments, were misclassified when including the data of the inter-annotation study. Together with the results from the inter-annotation study showing that the confused classes by the model are not exceptionally confused by the professional, it shows that the line between attempts from different errorclasses is very thin, and it depends on the dataset used in training which attempts are misclassified.

8.1.9 Including the data of the inter-annotation study

The data of the inter-annotation study was included in the trainingset to evaluate if this improved the performance of the models. The participants of this study were untrained adults, who were coached by an untrained professional. Their attempts were labeled by the professional as if they were children, as otherwise many attempts were labeled with label five, based on the fact that the duration was less than three seconds.

The attempts of the adults were labeled more often with multiple errors than the attempts from the children from the hospital experiments, with a difference of 35.8%. Additionally, the errors made during the attempts of the inter-annotation study are distributed over seven less errorclasses.

The addition of the data of the inter-annotation study decreased the performance of the best models trained on the trainingset without this data. When the models were trained and evaluated on only the data from the inter-annotation study, the performance was also lower, however the difference was smaller.

The decrease in performance implicates that a model performing well on a certain dataset, does not perform well on another dataset. This could be due to the creation of more noise in the dataset by using labels of two professionals when training on data of both studies. However, it could also be due to the mix of children and adults data. When training and evaluating on the adults dataset only, the performance was also lower. This could be due to this data being noisier, or because

of the data labeled by another professional results in another model being more appropriate for this dataset, or because the dataset consisted of data from adults. To exclude the option that the decrease in performance is due to mixing children and adults data, data from children should be collected and labeled by another professional to evaluate if this still decreases the performance. This was not possible for this study, due to the COVID-19 crisis.

8.1.10 Comparison with the rule-based approach

The best performing models based on machine learning are compared to the rule-based approach, explained in section 6.1.4. This approach bases its error detection on rules. The approach is designed to detect errors one (unsatisfactory start), three (an extra breath taken during the attempt), five (attempts that were terminated too early), six (no maximal effort), seven (cough), and glottis closure. Some rules represent multiple errors.

The rule-based approach outperformed the machine learning approach slightly with a difference in recall of 0.06. However, the rule-based approach only classified twenty-six out of 271 attempts right, compared to 166 by the machine learning approach. Besides, it misclassified all attempts from the zero errorclass (technically correct attempts). Thirdly, the rule-based approach is not distinctive as not meeting a rule may point to different errors. Finally, only six different errors can be detected. Thus, although the rule-based approach outperforms the machine learning approach based on the performance metrics, the machine learning approach is recommended, as it is a more extensive and accurate approach which will be more useful in providing feedback during home spirometry, as it is more distinctive, and is able to detect errors from more classes.

8.2 Inter-annotation study

We have seen that the inter-rater agreement is very low. Independent of the labelset used, the agreement, represented by the Cohen's kappa score, was found to be in a range of -0.123 to 0.380. This ranges from a negative agreement to a moderate agreement (refer to table 6.5 for the interpretation).The highest agreement was between rater two and three. These two raters are from the same hospital, and have the same background in training and experience. This shows that background makes a difference in how one looks at spirometry attempts.

The low inter-rater agreement score shows that it is not evident which errors are in spirometry attempts. Where rater one labeled 136 of the attempts as not containing an error, rater two and three labeled only eighty-six and fourty-five attempts as

such. Additionally, rater, one, two and three labeled respectively zero, fourty-eight, and fifty-seven attempts as containing multiple errors. These deviations shows that there is a big difference in the errors detected by professionals, and there is a thin line between whether an attempts is labeled as containing an error or not.

The intra-rater agreement of rater one, two, and three were respectively in a range of 0.862 to 0.865, which is seen as strong, 0.724 to 0.780, which is seen as moderate, and 0.648 to 0.860, which is seen as moderate to strong, for the three labelsets. The biggest deviations were in labeling an attempt as not containing an error, while it was labeled as containing an error in the other round, and labeling attempts as containing multiple errors, while these were labeled as containing only one error during the other round. This adds to the argument that there is a thin line between whether an attempt is labeled as containing an error or not.

The inter-rater agreement found in literature is much higher, with a lowest kappa score for the inter-rater agreement of 0.34 [54]. However, these raters were only asked to label if an attempt should be rejected or not. That is much more evident than assessing and comparing if one or multiple of twenty-two errors, giving tons of options, are in the attempts. An option to measure the inter-rater agreement is to give the raters a list of the most important errors, and ask the raters to only assess if these errors are in the attempts. This reduces the list of options and will probably result in a higher inter-rater agreement. However, this does not give a complete view of which errors are really in the attempts, as is done in this research.

The intra-rater agreements found in this research are lower than the intra-rater agreements found in literature, which are 98% to 99% [55]. One of the raters who assessed the attempts in this research stated in the communication with the researcher that they thought the errror detection in the data to be assessed was not straightforward, and that it was hard to determine if and which errors were in the attempts. This shows again the thin line between attempts with and without an error. To assess the intra-rater agreement more fairly, the same strategy as for the inter-rater agreement should be used; ask the raters to only allocate the most important errors, of which a list is provided by the researcher.

8.2.1 Implications for an error detection algorithm

The inter-rater agreements show that error detection in spirometry attempts is not evident. The algorithm proposed in this research uses the labels assigned to the attempts by a professional. However, if labels of another professional were used, the performance could vary heavily, based on the results of the inter-annotation study. This also means that a model trained on labels by one professional, can not be used by another professional as an attempt seen by the model as not containing an error

may not be seen this way by the new professional. This is one of the proposed reasons why the addition of the data from the inter-annotation study decreased the performance.

To be able to design an error detection algorithm, the first step is to sharpen the rules of what an error looks like, and when an attempt contains this error.

8.3 Comparison of coaching by a metaphor versus by a professional

The PEF, FEV_1, FVC values, and the number of errors in the attempts performed when the subject is coached by a professional amd by a metaphor were compared.

No significant differences were found. This shows that the quality of the attempts, represented by the PEF, FEV_1, FVC, and the number of errors, did not decrease significantly when the metaphors were used to coach the subject instead of a professional.

However, it is remarkable that the number of errors when coached by a professional are higher compared to when coached by the metaphor, as the coaching by the professional is expected to be more advanced. It could be due to a research bias, as the professional labeling the errors is part of the SpiroPlay project, and so not independent. Therefore, the absence of a significant difference in errors should be interpreted cautiously. The PEF, FEV_1, FVC values were calculated without the intervention of the professional, resulting in these values not being biased due to this research bias.

Literature shows that PEF values increase when the children play a game over being coached by a professional, while the FVC and FEV_1 values do not differ significantly [52]. This is different than the results found in this research. This could be due to the fact that the metaphors from literature focus primarily on PEF, while the metaphors in the present project also focus on reaching an acceptable FVC and FEV_1. Besides, it could be the case that the coaching by a professional in this project is different than the coaching of the professional in the study of Gracchi et al. (2003) [52].

The results in this research show that the metaphors used in the SpiroPlay system can be used to coach the children during home spirometry attempts, without a significant decrease in quality, based on FVC, FEV_1, PEF values, and presumably the number of errors.

8.4 Applicability of the system in home spirometry

The goal of the SpiroPlay system is to improve home monitoring of asthma, and to make it possible to monitor asthma patients more frequently, by delivering feedback to the child based on the errors that occured during the test, and by providing metaphors to steer the blowing behaviour of the child.

The error detection approach proposed in this study was based on a dataset with fifteen different errors, of which seven were based on the criteria which have to be met for a technically correct attempt, according to Miller et al. (2005) [3]. The dataset did not contain attempts with errors belonging to the criterias that maximal inhalation is needed for a technically correct attempt, and that no glottis closure and hesitation during the attempt is allowed. Therefore, the model could not be trained and evaluated on these errors.

The best performing model for the binary labelset was able to determine with a recall of 0.864, and a precision at 100% recall of 0.678 if an attempt is performed technically correct, based on the errors present in the dataset. This is adequate enough to be used in a real-life system. The performance of the best performing models for the labelsets 'combined' and 'all' is too low to be used in a real life system.

Although the binary model is accurate enough to be used in a real life system, the negative to minimal inter-rater agreements imply that a model trained on labels assigned by one professional, will not be useful for another professional. Therefore, before this algorithm can be used in a real life situation, the rules of what an error looks like, and when an attempt contains this error, should be sharpened. Besides, to be able to base the decision if an attempt is technically correct or not on the whole criteria list by Miller et al. (2005) [3], attempts where the subject did not inhale maximally, where glottis closure, and hesitation during the attempt occurred should be included in the dataset.

The second goal of the system SpiroPlay is to offer metaphors to steer the blowing behaviour of the child during home spirometry. In the hospital, the child is coached by a professional. However, this professional is not available at home. Therefore, another way of coaching should be designed. The present study has shown that metaphors are a good approach to coach the children at home, as no significant difference was found in quality defined by the FVC, FEV_1, and PEF values, and presumably the number of errors, when comparing spirometry attempts coached by a professional and by a metaphor.

In summary, the metaphors can be used as a coaching manner in the system SpiroPlay. However, the error detection approach cannot be designed and used in this system before the rules of what an error looks like are sharpened.

8.5 Scientific contributions

This study focused on three goals; designing and evaluating an error detection approach based on machine learning, the evaluation of the agreement in detecting errors in spirometry data by multiple professionals, and the evaluation of the effect of the metaphors on the quality of spirometry measurements.

This study showed that the agreement in detecting errors by professionals is very low, meaning that before a generic error detection algorithm can be designed which will support all professionals in monitoring asthma, stricter guidelines are needed to assess when an error occurred. Therefore, a first step is to generate stricter rules.

Although this study revealed that automatic error detection based on machine learning is not yet feasible, it also established that using metaphors as coaching method did not result in significant quality loss when compared to the coaching method used by the professional. Therefore, using these metaphors during home spirometry will help the children to blow good attempts.

The remained quality of the spirometry attempts using metaphors as coaching method brings the medical world a step closer to home monitoring of asthma, which will result in lower healthcare costs, and improved freedom for the children as they do not have to visit the hospital as often. However, as an error detection approach supporting all professionals is not achievable yet, the attempts still have to be assessed manually by the professional to select the three best attempts, until stricter rules to determine when an error occurred are introduced.

8.6 Strengths and limitations

This broad research focused on multiple topics; designing and evaluating an error detection approach based on machine learning, an inter-annotation study was designed and performed to assess the inter- and intra-rater agreements between multiple professionals detecting errors in spirometry data, and the question if the difference in quality of a spirometry attempt was significant when the child is coached by a professional and when coached by a metaphor during a spirometry attempt was answered. The processes to answer the questions belonging to these three areas of focus where executed extensively and well-considered. An example of the extensiveness is the use of different labelsets, different featuresets, different models and hyperparameters, different balancing techniques, and different machine learning techniques to find the best possible models. An example of the level of consideration is the execution of a test recording before the actual inter-annotation study, to be certain that the right angels were used when recording the subjects.

However, this research also contains limitations. First, this research states that

the rules to determine if an error occurred or not are not strict enough to design a generic error detection algorithm. This statement is based on the low inter-rater agreement, and on the decrease in performance of the models when trained on (a mix of children and) adults data. We expect the decrease in performance to be mainly due to the professionals labeling data differently. However, the effect of using data from adults instead of children could not be excluded as the inter-annotation study was performed with adults. Although performing this study with children was not possible due to the COVID-19 crisis, this is a limitation of this research.

Secondly, the data the error detection algorithm was based on did not cover all criteria which should be met, according to Miller et al. (2005) [3], for a technically correct attempt, resulting in an incomplete decision if an attempt is technically correct or not when using the models based on this dataset. Data covering all criteria was not available during this research, but should be collected for future research.

Thirdly, although literature states that compliance is an issue in home spirometry attempts, this research did not investigate if this is also the case when using the SpiroPlay system. This was out of the scope of this research, but should be looked into during future research.

Finally, the criteria by Miller et al. (2005) [3] were updated during this study. The main difference is in the more strict definition of the End of Test, now called the 'End of Forced Expiration', in which there is more focus on a full inhalation after expiration. This part of the manoevre is not taken into account in this research. Therefore, the new criteria described by Miller et al. (2019) [67] should be used during future research.

Chapter 9

Conclusion

This chapter concludes this research by answering the research questions, asked in chapter 5.

9.1 Error detection

The first goal of this research was to design and evaluate an error detection approach using machine learning techniques to detect errors in spirometry data. Different classifiers, featuresets, labelsets, and balancing techniques were evaluated. Also, stacking was used as an ensemble method, and a proposed decision tree was evaluated.

The best performing models for all labelsets were the SVMs, using the spirometry parameters as input for the labelsets 'combined' and 'all', and the filtered featureset for the binary labelset. Performance increased after balancing, and ROS outperformed SMOTE as a balancing technique.

The recall and precision at 100% recall of the best model for the binary labelset were 0.864 and 0.678. According to a professional, this is accurate enough to be used in a real life system.

The recall and precision at 100% recall of the best models when using the combined labelset were 0.525 and 0.134, and when using the labelset 'all' 0.322 and 0.063. According to the professional, these performance are too poor to use in a real life system.

Stacking and the proposed decision tree did not outperform the best single models.

When comparing the machine learning approach to the rule-based approach, we conclude that the rule-based approach outperforms the machine learning approach based on precision, recall, and the F1-score. However, the rule-based approach classified only twenty-six out of 271 attempts right compared to 166 by the machine

learning approach, is only designed to classify six different errors, instead of the fifteen the machine learning approach is trained on, and is not distinctive as not meeting a rule can point to multiple errors. Therefore, the machine learning approach is recommended over the rule-based approach.

9.2 Inter-annotation study

The second goal of this research was to assess the agreement in error detection by professionals. To meet this goal, an inter-annotation study was performed. Thirteen adults participated in this study, and the attempts were labeled by three professionals. The agreement was calculated for the three labelsets also used in the designing of the error detection algorithm.

The Cohen's kappa score representing the inter-rater agreement ranged from -0.123 to 0.380, which can be interpreted as a negative to minimal agreement. The intra-rater agreement of rater one for the three labelsets ranged from 0.862 to 0.865, which is a strong agreement, of rater two from 0.724 to 0.780, which is a moderate agreement, and for rater three from 0.648 to 0.860, which is a moderate to strong agreement. Especially the low inter-rater agreements reveal that professionals assess the spirometry attempts very differently. When connecting these results to the designing of the error detection algorithm based on machine learning, we conclude that a model based on labels of one professional does not support another professional when used in a system for home monitoring of asthma. Therefore, before a generic error detection algorithm can be designed, the rules describing the different errors should first be sharpened.

9.3 Comparison of coaching by a metaphor versus by a professional

The third goal of this research was to evaluate if there is a difference in quality of spirometry attempts coached by a professional and by a metaphor. This quality was represented by the PEF, FEV_1, FVC values, and the number of errors made during the attempts, based on the labeling of the professional. The data from the hospital experiments was used to evaluate this difference.

The results show that there is no significant difference in PEF, FEV_1, FVC values, and the number of errors. However, there could have been a research bias when labeling the errors in the spirometry attempts. Therefore, the absence of a significant difference in the number of errors should be taken cautiously.

These results demonstrate that using the metaphors during home spirometry is a good way to coach the children during home spirometry attempts, without significant loss of quality based on PEF, FEV_1, FVC, and presumably the number of errors.

9.4 Final remarks

The findings of this research implicate that it is possible to perform home spirometry, coached by metaphors, without quality loss based on the measured PEF, FEV_1, FVC values, and presumably the number of errors.

The evaluation of the error detection algorithm show that it is possible to design an algorithm which determines if an attempt is technically correct or not. However, the negative to minimal inter-rater agreements show that professionals detect different errors in the same spirometry attempts. Therefore, before a generic error detection algorithm can be designed and used in a real life system, stricter rules should be introduced describing if and which error occurred during an attempt.

Bibliography

[1] A. Felman, “Asthma: Definition, types, causes, and diagnosis,” Nov 2018. [Online]. Available: https://www.medicalnewstoday.com/articles/323523.php

[2] M. van der Kamp, “Wearcon : Wearable home-monitoring in asthmatic children.” August 2017. [Online]. Available: http://essay.utwente.nl/73924/

[3] M. R. Miller, J. Hankinson, V. Brusasco, F. Burgos, R. Casaburi, A. Coates, R. Crapo, P. Enright, C. P. M. van der Grinten, P. Gustafsson, R. Jensen, D. C. Johnson, N. MacIntyre, R. McKay, D. Navajas, O. F. Pedersen, R. Pellegrino, G. Viegi, and J. Wanger, “Standardisation of spirometry,” *European Respiratory Journal*, vol. 26, no. 2, pp. 319–338, 2005, doi: 10.1183/09031936.05.00034805. [Online]. Available: https://erj.ersjournals.com/content/26/2/319

[4] WebMD, “Asthma: Causes, symptoms, diagnosis, treatment,” 2019. [Online]. Available: https://www.webmd.com/asthma/what-is-asthma#4

[5] R. Delden, M. van der Kamp, A. Moreno, F. Sieverink, B. Thio, R. Klaassen, M. Gorrissen, and R. Stam, “Steering into the flow - gamified spirometry for telemonitoring children with asthma,” December 2018, Research Proposal, [Online]. On demand: https://www.utwente.nl/en/techmed/innovation/funds-vouchers/pioneers-in-healthcare/projects-archief/projects2018/.

[6] WebMD, “Asthma: Causes, symptoms, diagnosis, treatment,” 2019. [Online]. Available: https://www.webmd.com/asthma/what-is-asthma#3

[7] WebMD, “Asthma: Causes, symptoms, diagnosis, treatment,” 2019. [Online]. Available: https://www.webmd.com/asthma/what-is-asthma#1

[8] WebMD, “Asthma: Causes, symptoms, diagnosis, treatment,” 2019. [Online]. Available: https://www.webmd.com/asthma/what-is-asthma#2

[9] Mayo Clinic, “Asthma,” Sep 2018. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/asthma/symptoms-causes/syc-20369653

[10] P. Brand and R. Roorda, “Usefulness of monitoring lung function in asthma,” *Archives of disease in childhood*, vol. 88, pp. 1021–5, 12 2003, doi: 10.1136/adc.88.11.1021. [Online]. Available: https://doi.org/10.1136/adc.88.11.1021

[11] GLI, “Global lung function initiative,” 2019. [Online]. Available: https://www.ers-education.org/guidelines/global-lung-function-initiative.aspx

[12] “NuvoAir AB,” 2018. [Online]. Available: https://www.nuvoair.com/

[13] L. W. Waldemar Tomalak W., Radliski J., “Quality of spirometric measurements in children younger than 10 years of age in the light of the reccommendations,” *Advances in Respiratory Medicine*, vol. 76, no. 6, pp. 421–425, 2008, doi: 10.1002/ppul.20908. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ppul.20908

[14] J. S. Loeb, W. C. Blower, J. F. Feldstein, B. A. Koch, A. L. Munlin, and W. D. Hardie, “Acceptability and repeatability of spirometry in children using updated ATS/ERS criteria,” *Pediatric Pulmonology*, vol. 43, no. 10, pp. 1020–1024, 2008, doi: 10.1002/ppul.20908. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ppul.20908

[15] R. Thompson, R. J. Delfino, T. Tjoa, E. Nussbaum, and D. Cooper, “Evaluation of daily home spirometry for school children with asthma: New insights,” *Pediatric Pulmonology*, vol. 41, no. 9, pp. 819–828, 2006, doi: 10.1002/ppul.20449. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ppul.20449

[16] BTS/ARTP Liaison Committee *et al.*, “Guidelines for the measurement of respiratory function. recommendations of the British Thoracic Society and the Association of Respiratory Technicians and Physiologists,” *Respir Med*, vol. 88, pp. 165–194, 1994, doi: 10.1016/S0954-6111(05)80346-4.

[17] N. Beydon, S. D. Davis, E. Lombardi, J. L. Allen, H. G. M. Arets, P. Aurora, H. Bisgaard, G. M. Davis, F. M. Ducharme, H. Eigen, M. Gappa, C. Gaultier, P. M. Gustafsson, G. L. Hall, Z. Hantos, M. J. R. Healy, M. H. Jones, B. Klug, K. C. Ldrup Carlsen, S. A. McKenzie, F. Marchal, O. H. Mayer, P. J. F. M. Merkus, M. G. Morris, E. Oostveen, J. J. Pillow, P. C. Seddon, M. Silverman, P. D. Sly, J. Stocks, R. S. Tepper, D. Vilozni, and N. M. Wilson, “An official american thoracic society/european respiratory society statement: Pulmonary function testing in preschool children,” *American Journal of Respiratory and Critical Care Medicine*, vol. 175, no. 12, pp. 1304–1345, 2007, pMID: 17545458. [Online]. Available: https://doi.org/10.1164/rccm.200605-642ST

[18] H. Reddel, S. Ware, C. Salome, C. Jenkins, and A. Woolcock, “Pitfalls in processing home electronic spirometric data in asthma,” *European Respiratory Journal*, vol. 12, no. 4, pp. 853–858, 1998, doi: 10.1183/09031936.98.12040853. [Online]. Available: https://erj.ersjournals.com/content/12/4/853

[19] Mayo Clinic, “Budesonide (inhalation route) description and brand names,” Oct 2019. [Online]. Available: https://www.mayoclinic.org/drugs-supplements/budesonide-inhalation-route/description/drg-20071233

[20] P. Gannon, J. Belcher, C. Pantin, and P. Burge, “The effect of patient technique and training on the accuracy of self-recorded peak expiratory flow,” *European Respiratory Journal*, vol. 14, no. 1, pp. 28–31, 1999, doi: 10.1034/j.1399-3003.1999.14a07.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1399-3003.1999.14a07.x

[21] H. Arets, H. Brackel, and C. van der Ent, “Forced expiratory manoeuvres in children: do they meet ATS and ERS criteria for spirometry?” *European Respiratory Journal*, vol. 18, no. 4, pp. 655–660, 2001, doi: 10.1183/09031936.01.00204301. [Online]. Available: https://erj.ersjournals.com/content/18/4/655

[22] K. Desmond, P. Allen, D. Demizio, T. Kovesi, and A. Coates, “Redefining End of Test (EOT) criteria for pulmonary function testing in children,” *American Journal of Respiratory and Critical Care Medicine*, vol. 156, no. 2, pp. 542–545, 1997, doi: 10.1164/ajrccm.156.2.9610116. [Online]. Available: https://doi.org/10.1164/ajrccm.156.2.9610116

[23] K. M. Mortimer, A. Fallot, J. R. Balmes, and I. B. Tager, “Evaluating the use of a portable spirometer in a study of pediatric asthma,” *Chest*, vol. 123, no. 6, pp. 1899 – 1907, 2003, doi: 10.1378/chest.123.6.1899. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0012369216348097

[24] P. J. Chowienczyk, D. H. Parkin, C. P. Lawson, and G. M. Cochrane, “Do asthmatic patients correctly record home spirometry measurements?” *BMJ*, vol. 309, no. 6969, p. 1618, 1994, doi: 10.1136/bmj.309.6969.1618. [Online]. Available: https://www.bmj.com/content/309/6969/1618

[25] D. C. Wensley and M. Silverman, “The quality of home spirometry in school children with asthma,” *Thorax*, vol. 56, no. 3, pp. 183–185, 2001, doi: 10.1136/thorax.56.3.183. [Online]. Available: https://thorax.bmj.com/content/56/3/183

[26] A. S. Pelkonen, K. Nikander, and M. Turpeinen, “Reproducibility of home spirometry in children with newly diagnosed asthma,” *Pediatric Pulmonology*, vol. 29, no. 1, pp. 34–38, 2000, doi: 10.1002/(SICI)1099-0496(200001)29:1¡34::AID-PPUL6¿3.0.CO;2-O. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-0496%28200001%2929%3A1%3C34%3A%3AAID-PPUL6%3E3.0.CO%3B2-O

[27] S. Redline, E. C. Wright, M. Kattan, C. Kercsmar, and K. Weiss, “Short-term compliance with peak flow monitoring: Results from a study of inner city children with asthma,” *Pediatric Pulmonology*, vol. 21, no. 4, pp. 203–210, 1996, doi: 10.1002/(SICI)1099-0496(199604)21:4¡203::AID-PPUL1¿3.0.CO;2-P. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-0496%28199604%2921%3A4%3C203%3A%3AAID-PPUL1%3E3.0.CO%3B2-P

[28] M. E. Hyland, C. A. Kenyon, R. Allen, and P. Howarth, “Diary keeping in asthma: comparison of written and electronic methods.” *BMJ*, vol. 306, no. 6876, pp. 487–489, 1993, doi: 10.1136/bmj.306.6876.487. [Online]. Available: https://www.bmj.com/content/306/6876/487

[29] P. Verschelden, A. Cartier, J. L’Archeveque, C. Trudeau, and J. Malo, “Compliance with and accuracy of daily self-assessment of peak expiratory flows (PEF) in asthmatic subjects over a three month period,” *European Respiratory Journal*, vol. 9, no. 5, pp. 880–885, 1996, doi: 10.1183/09031936.96.09050880. [Online]. Available: https://erj.ersjournals.com/content/9/5/880

[30] A. F. J. Brouwer, R. J. Roorda, and P. L. P. Brand, “Home spirometry and asthma severity in children,” *European Respiratory Journal*, vol. 28, no. 6, pp. 1131–1137, 2006, doi: 10.1183/09031936.06.00118205. [Online]. Available: https://erj.ersjournals.com/content/28/6/1131

[31] P. D. Sly and F. Flack, “Is home monitoring of lung function worthwhile for children with asthma?” *Thorax*, vol. 56, no. 3, pp. 164–165, 2001, doi: 10.1136/thorax.56.3.164. [Online]. Available: https://thorax.bmj.com/content/56/3/164

[32] P. Brand, E. Duiverman, D. Postma, H. Waalkens, K. Kerrebijn, and E. van Essen-Zandvliet, “Peak flow variation in childhood asthma: relationship to symptoms, atopy, airways obstruction and hyperresponsiveness. Dutch CNSLD Study Group,” *European Respiratory Journal*, vol. 10, no. 6, pp. 1242–1247, 1997, doi: 10.1183/09031936.97.10061242. [Online]. Available: https://erj.ersjournals.com/content/10/6/1242

[33] P. L. P. Brand, E. J. Duiverman, H. J. Waalkens, E. E. M. van Essen-Zandvliet, K. F. Kerrebijn, and , "Peak flow variation in childhood asthma: correlation with symptoms, airways obstruction, and hyperresponsiveness during long term treatment with inhaled corticosteroids," *Thorax*, vol. 54, no. 2, pp. 103–107, 1999, doi: 10.1136/thx.54.2.103. [Online]. Available: https://thorax.bmj.com/content/54/2/103

[34] J. E. Gern, P. A. Eggleston, K. C. Schuberth, N. Eney, E. O. Goldstein, M. E. Weiss, and N. Adkinson, "Peak flow variation in childhood asthma: A three-year analysis," *Journal of Allergy and Clinical Immunology*, vol. 93, no. 4, pp. 706 – 716, 1994, doi: 10.1016/0091-6749(94)90250-X. [Online]. Available: http://www.sciencedirect.com/science/article/pii/009167499490250X

[35] A. C. Ferguson, "Persisting airway obstruction in asymptomatic children with asthma with normal peak expiratory flow rates," *Journal of Allergy and Clinical Immunology*, vol. 82, no. 1, pp. 19 – 22, 1988, doi: 10.1016/0091-6749(88)90045-0. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0091674988900450

[36] I. Charlton, G. Charlton, J. Broomfield, and M. A. Mullee, "Evaluation of peak flow and symptoms only self management plans for control of asthma in general practice." *BMJ*, vol. 301, no. 6765, pp. 1355–1359, 1990, doi: 10.1136/bmj.301.6765.1355. [Online]. Available: https://www.bmj.com/content/301/6765/1355

[37] N. Drummond, M. Abdalla, J. A. G. Beattie, J. K. Buckingham, T. Lindsay, L. M. Osman, S. J. Ross, A. Roy-Chaudhury, I. Russell, M. Turner, J. A. R. Friend, J. S. Legge, and J. G. Douglas, "Effectiveness of routine self monitoring of peak flow in patients with asthma," *BMJ*, vol. 308, no. 6928, pp. 564–567, 1994, doi: 10.1136/bmj.308.6928.564. [Online]. Available: https://www.bmj.com/content/308/6928/564

[38] D. Wensley and M. Silverman, "Peak flow monitoring for guided self-management in childhood asthma," *American Journal of Respiratory and Critical Care Medicine*, vol. 170, no. 6, pp. 606–612, 2004, doi: 10.1164/rccm.200307-1025OC. [Online]. Available: https://doi.org/10.1164/rccm.200307-1025OC

[39] L. Agertoft and S. Pedersen, "Effect of long-term treatment with inhaled budesonide on adult height in children with asthma," *New England Journal of Medicine*, vol. 343, no. 15, pp. 1064–1069, 2000, doi:

10.1056/NEJM200010123431502. [Online]. Available: https://doi.org/10.1056/NEJM200010123431502

[40] The Childhood Asthma Management Program Research Group, "Long-term effects of budesonide or nedocromil in children with asthma," *New England Journal of Medicine*, vol. 343, no. 15, pp. 1054–1063, 2000, doi: 10.1056/NEJM200010123431501. [Online]. Available: https://doi.org/10.1056/NEJM200010123431501

[41] C. D. Ramsey, J. C. Celedn, D. L. Sredl, S. T. Weiss, and M. M. Cloutier, "Predictors of disease severity in children with asthma in Hartford, Connecticut," *Pediatric Pulmonology*, vol. 39, no. 3, pp. 268–275, 2005, doi: 10.1002/ppul.20177. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ppul.20177

[42] J. Hankinson, J. Odencrantz, and K. Fedan, "Spirometric reference values from a sample of the general u.s. population," *American Journal of Respiratory and Critical Care Medicine*, vol. 159, no. 1, p. 179187, 1999, doi: 10.1164/ajrccm.159.1.9712108.

[43] A. L. Fuhlbrigge, B. T. Kitch, A. Paltiel, K. M. Kuntz, P. J. Neumann, D. W. Dockery, and S. T. Weiss, "FEV1 is associated with risk of asthma attacks in a pediatric population," *Journal of Allergy and Clinical Immunology*, vol. 107, no. 1, pp. 61 – 67, 2001, doi: 10.1067/mai.2001.111590. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0091674901183053

[44] J. D. Spahn, R. Cherniack, K. Paull, and E. W. Gelfand, "Is forced expiratory volume in one second the best measure of severity in childhood asthma?" *American Journal of Respiratory and Critical Care Medicine*, vol. 169, no. 7, pp. 784–786, 2004, doi: 10.1164/rccm.200309-1234OE. [Online]. Available: https://doi.org/10.1164/rccm.200309-1234OE

[45] M. R. Simon, V. M. Chinchilli, B. R. Phillips, C. A. Sorkness, R. F. Lemanske, S. J. Szefler, L. Taussig, L. B. Bacharier, and W. Morgan, "Forced expiratory flow between 25% and 75% of vital capacity and FEV1/forced vital capacity ratio in relation to clinical and physiological parameters in asthmatic children with normal fev1 values," *Journal of Allergy and Clinical Immunology*, vol. 126, no. 3, pp. 527 – 534.e8, 2010, doi: 10.1016/j.jaci.2010.05.016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0091674910008171

[46] L. B. Bacharier, R. C. Strunk, D. Mauger, D. White, R. F. Lemanske, and C. A. Sorkness, "Classifying asthma severity in children," *American*

Journal of Respiratory and Critical Care Medicine, vol. 170, no. 4, pp. 426–432, 2004, doi: 10.1164/rccm.200308-1178OC. [Online]. Available: https://doi.org/10.1164/rccm.200308-1178OC

[47] V. Ratageri, S. Kabra, R. Lodha, S. Dwivedi, and V. Seth, "Brief report. Lung function tests in asthma: which indices are better for assessment of severity?" *Journal of Tropical Pediatrics*, vol. 47, no. 1, pp. 57–59, 02 2001, doi: 10.1093/tropej/47.1.57. [Online]. Available: https://doi.org/10.1093/tropej/47.1.57

[48] J. O. Warner, "Asthma: a follow up statement from an international paediatric asthma consensus group." *Archives of Disease in Childhood*, vol. 67, no. 2, p. 240248, Jan 1992.

[49] D. Vilozni, M. Barker, H. Jellouschek, G. Heimann, and H. Blau, "An interactive computer-animated system (SpiroGame) facilitates spirometry in preschool children," *American Journal of Respiratory and Critical Care Medicine*, vol. 164, no. 12, pp. 2200–2205, 2001, doi: 10.1164/ajrccm.164.12.2101002. [Online]. Available: https://doi.org/10.1164/ajrccm.164.12.2101002

[50] D. Vilozni, A. Barak, O. Efrati, A. Augarten, C. Springer, Y. Yahav, and L. Bentur, "The role of computer games in measuring spirometry in healthy and asthmatic preschool children," *Chest*, vol. 128, no. 3, pp. 1146 – 1155, 2005, doi: 10.1378/chest.128.3.1146. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0012369215521302

[51] H. Eigen, H. Bieler, D. Grant, K. Christoph, D. Terrill, D. Heilman, W. Ambrosius, and R. Tepper, "Spirometric pulmonary function in healthy preschool children," *American Journal of Respiratory and Critical Care Medicine*, vol. 163, no. 3, pp. 619 – 623, 2001, doi: 10.1164/ajrccm.163.3.2002054.

[52] V. Gracchi, M. Boel, J. van der Laag, and C. van der Ent, "Spirometry in young children: should computer-animation programs be used during testing?" *European Respiratory Journal*, vol. 21, no. 5, pp. 872–875, 2003, doi: 10.1183/09031936.03.00059902. [Online]. Available: https://erj.ersjournals.com/content/21/5/872

[53] W. Kozlowska, P. Aurora, and J. Stocks, "The use of computer-animation programs during spirometry in preschool children," *European Respiratory Journal*, vol. 23, no. 3, pp. 494–495, 2004, doi: 10.1183/09031936.04.00126904. [Online]. Available: https://erj.ersjournals.com/content/23/3/494.2

[54] F. Velickovski, L. Ceccaroni, R. Marti, F. Burgos, C. Gistau, X. Alsina-Restoy, and J. Roca, "Automated spirometry quality assurance: Supervised learning from multiple experts," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 276–284, 2018, doi: 10.1109/JBHI.2017.2713988. [Online]. Available: https://doi.org/10.1109/JBHI.2017.2713988

[55] L. E. Tuomisto, V. Jarvinen, J. Laitinen, M. Erhola, M. Kaila, and P. E. Brander, "Asthma programme in finland: the quality of primary care spirometry is good," *Primary Care Respiratory Journal*, vol. 17, no. 4, p. 226231, Feb 2008, doi: 10.3132/pcrj.2008.00053. [Online]. Available: https://doi.org/10.3132/pcrj.2008.00053

[56] S. M. Seyedmehdi, M. Attarchi, T. Yazdanparast, and M. M. Lakeh, "Quality of spirometry tests and pulmonary function changes among industrial company workers in iran: a two-year before-and-after study following an intensive training intervention," *Primary Care Respiratory Journal*, vol. 22, no. 1, p. 8691, 2013, doi: 10.4104/pcrj.2013.00018. [Online]. Available: https://dx.doi.org/10.4104%2Fpcrj.2013.00018

[57] A. Z. Luo, E. Whitmire, J. W. Stout, D. Martenson, and S. Patel, "Automatic characterization of user errors in spirometry," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2017, pp. 4239–4242, doi: 10.1109/EMBC.2017.8037792.

[58] D. Sahin, E. D. Übeyli, G. Ilbay, M. Sahin, and A. B. Yasar, "Diagnosis of airway obstruction or restrictive spirometric patterns by multiclass Support Vector Machines," *Journal of Medical Systems*, vol. 34, no. 5, pp. 967–973, Oct 2010, doi: 10.1007/s10916-009-9312-7. [Online]. Available: https://doi.org/10.1007/s10916-009-9312-7

[59] P. Bright, M. Miller, J. Franklyn, and M. Sheppard, "The use of a neural network to detect upper airway obstruction caused by Goiter," *American Journal of Respiratory and Critical Care Medicine*, vol. 157, no. 6, pp. 1885–1891, 1998, doi: 10.1164/ajrccm.157.6.9705022. [Online]. Available: https://doi.org/10.1164/ajrccm.157.6.9705022

[60] S. Manoharan, M. Veezhinathan, and S. Ramakrishnan, "Comparison of two ann methods for classification of spirometer data," *Measurement Science Review*, vol. 8, no. 3, pp. 53–57, Jan 2008, doi: 10.2478/v10048-008-0014-y.

[61] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," *CoRR*, vol. abs/1610.07717, 2016. [Online]. Available: http://arxiv.org/abs/1610.07717

[62] J. S. Felix A. Gers, Nicol N. Schraudolph, "Learning precise timing with LSTM recurrent networks," *Journal of Machine Learning Research*, vol. 3, pp. 115–143, Aug 2002. [Online]. Available: http://www.jmlr.org/papers/v3/gers02a.html

[63] S. Sun, "Meta-analysis of Cohen's kappa," *Health Services and Outcomes Research Methodology*, vol. 11, no. 3, pp. 145–163, Dec 2011, doi:10.1007/s10742-011-0077-3. [Online]. Available: https://doi.org/10.1007/s10742-011-0077-3

[64] M. L. Mchugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, p. 276282, 2012, doi: 10.11613/bm.2012.031.

[65] R. M. Tischio and G. M. Weiss, "Identifying classification algorithms most suitable for imbalanced data," 2019.

[66] Z. Jaadi, "Eveything you need to know about interpreting correlations," Oct 2019. [Online]. Available: https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8

[67] B. L. Graham, I. Steenbruggen, M. R. Miller, I. Z. Barjaktarevic, B. G. Cooper, G. L. Hall, T. S. Hallstrand, D. A. Kaminsky, K. McCarthy, M. C. McCormack, C. E. Oropez, M. Rosenfeld, S. Stanojevic, M. P. Swanney, and B. R. Thompson, "Standardization of spirometry 2019 update. an official american thoracic society and european respiratory society technical statement," *American Journal of Respiratory and Critical Care Medicine*, vol. 200, no. 8, pp. e70–e88, 2019, doi:10.1164/rccm.201908-1590ST. [Online]. Available: https://doi.org/10.1164/rccm.201908-1590ST

[68] M. W. Watkins and M. Pacheco, "Interobserver agreement in behavioral research: Importance and calculation," *Journal of Behavioral Education*, vol. 10, no. 4, pp. 205–212, Dec 2000, doi: 10.1023/A:1012295615144. [Online]. Available: https://doi.org/10.1023/A:1012295615144

[69] M. Nguyen, "Illustrated guide to lstm's and gru's: A step by step explanation," Jul 2019. [Online]. Available: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[70] T. Yiu, "Understanding neural networks," Aug 2019. [Online]. Available: https://towardsdatascience.com/understanding-neural-networks-19020b758230

[71] C. McCormick, "Radial basis function network (rbfn) tutorial," Aug 2013. [Online]. Available: https://mccormickml.com/2013/08/15/radial-basis-function-network-rbfn-tutorial/

[72] C. Maklin, “Adaboost classifier example in python,” Jul 2019. [Online]. Available: https://towardsdatascience.com/machine-learning-part-17-boosting-algorithms-adaboost-in-python-d00faac6c464

[73] R. Gandhi, “Support vector machine - introduction to machine learning algorithms,” Jul 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[74] “Precision-recall,” 2019. [Online]. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html#:~:text=Theprecision-recallcurveshows,alowfalsenegativerate.

[75] Stephanie, “Q Q plots: Simple definition example,” Aug 2019. [Online]. Available: https://www.statisticshowto.datasciencecentral.com/q-q-plots/

[76] M. Christ, N. Braun, and J. Neuffer, “Overview on extracted features,” 2020. [Online]. Available: https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html

Appendix A

Background information method

A.1 Kappa score

The Cohens kappa score is calculated using the following formula:

$$\kappa = \frac{PR(a) - PR(e)}{1 - PR(e)} \tag{A.1}$$

Pr(a) represents the actual observed agreement, Pr(e) represents the chance agreement.

The strength of this method is that it takes into consideration the chance that the professionals guess the label instead of knowing it.

A kappa score below 0.40 is seen as poor agreement, between 0.40 and 0.60 as fair, 0.60 to 0.75 as good, and above 0.75 as excellent agreement [68] .

Below a calculation example is given.

The formulas we need are:

$$\kappa = \frac{PR(a) - PR(e)}{1 - PR(e)} \tag{A.2}$$

$$PR(a) = \frac{Agreements}{Agreements + Disagreements} \tag{A.3}$$

$$PR(e) = \frac{X_1 * Y_1}{N^2} + \frac{X_2 * Y_2}{N^2} \tag{A.4}$$

Example data:

Observer 2 \ Observer 1	Yes	No
Yes	20	5
No	15	10

Table A.1: Example data for calculation Kappa score

Using the formulas A.2, A.3, and A.4 and the example data yields:

$$PR(a) = \frac{30}{50} = 0.6 \tag{A.5}$$

$$PR(e) = \frac{(20+5)*(20+15)}{50^2} + \frac{(15+10)*(5+10)}{50^2} = \frac{875}{2500} + \frac{375}{2500} = 0.35 + 0.15 = 0.5 \tag{A.6}$$

$$\kappa = \frac{0.6 - 0.5}{1 - 0.5} = \frac{0.1}{0.5} = 0.2 \tag{A.7}$$

A.2 Normalization

Min-max normalization:

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{A.8}$$

Z-normalization:

$$z = \frac{x_i - \mu}{\sigma} \tag{A.9}$$

A.3 Smoothing

To smooth the data a Savitzky-Golay filter is used. This low-pass filter uses a polynomial fit using $2n + 1$ neighbouring points, which include the point to be smoothed. N is at least the order of the polynomial.

An example formula is presented in equation A.10, where a second order polynomial and seven datapoints are used to smooth a datapoint.

$$y_t = (-2x_{t-3} + 3x_{t-2} + 6x_{t-1} + 7x_t + 6x_{t+1} + 3x_{t+2} - 2x_{t+3})/21 \tag{A.10}$$

A.4 K-fold cross validation

In K-fold cross validation, the dataset is divided in K subsamples, one used for testing and the rest for training. This is shown in Figure A.1. The model is trained K times. Every time another subsample is the test sample. The results are averaged to get an overall result which can be used to analyze the quality of the model.

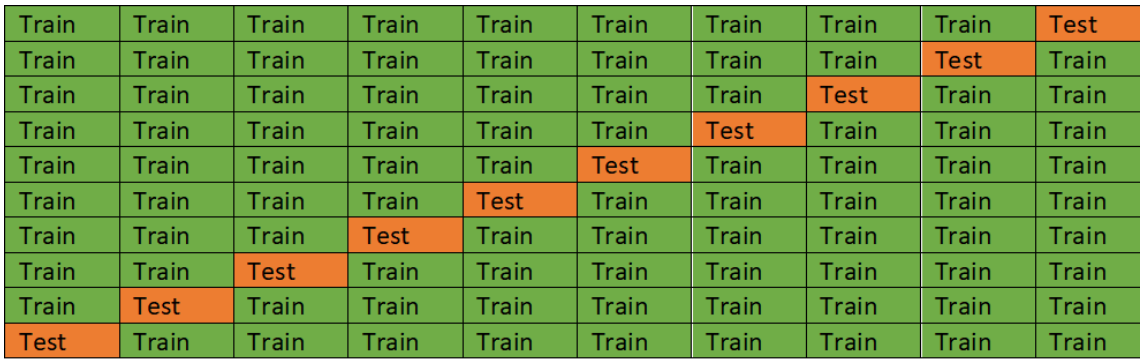


Figure A.1: K-fold cross validation

A.5 LSTM (Recurrent Neural Network)

Long-Short Term Memory networks (LSTMs) are trained using back-propagation. It makes use of three gates as shown in Figure A.2; the forget gate, the input gate, and the output gate. Using these gates, the model is able to learn which data is important to keep passing the relevant information down the chain of sequences.

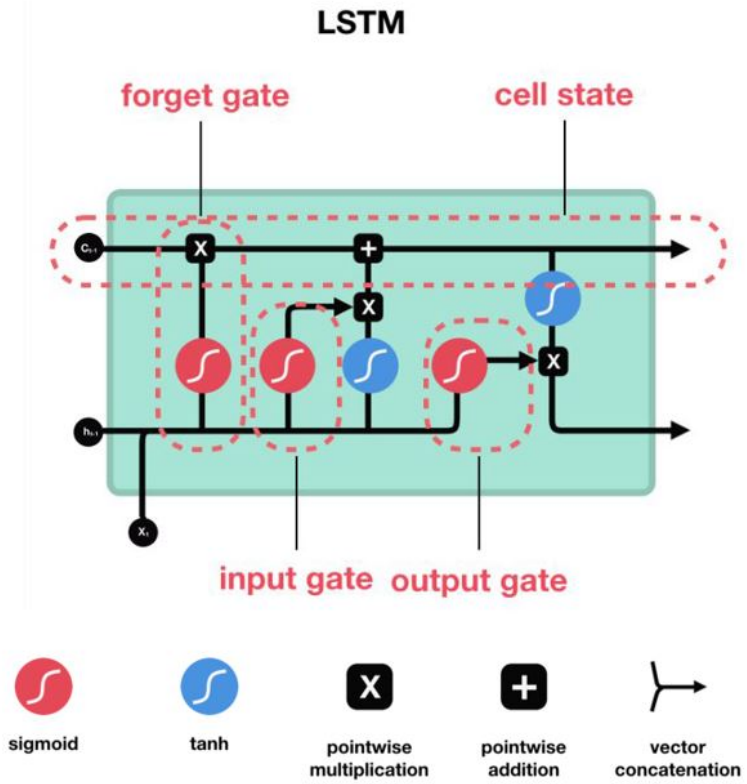


Figure A.2: Overview of an LSTM cell. Source: adapted from figures 1 and 12 of Nguyen (2019) [69]

The first gate is the forget state; this state determines which information should be kept by passing the information from the previous hidden state (h_{t-1} in figure A.2) and the current input (X_t in figure A.2) through a sigmoid function. This turns values between zero and one; the values closer to one are the ones that should be kept.

The input gate starts with passing the previous hidden state (h_{t-1} in figure A.2) and the current input (X_t in figure A.2) through a sigmoid function which transforms the values to values between zero and one to decide which values will be updated; zero means that the value is not important, one means it is important. The current input and the hidden state are also passed through a tanh function to transform the values to be between minus one and one as this is necessary for this network. The two outputs are multiplied; the output from the sigmoid function decides which information has to be kept from the output of the tanh function.

The output from the forget state and the previous cell state (c_{t-1} in figure A.2) are multiplied to drop values in the previous cell state, when these get multiplied by values from the forget gate which are near zero. This output is added pointwise with the output from the input gate, which creates the new cell state.

The output gate determines the new hidden state. First, the previous hidden state together with the current input is passed through a sigmoid function. Secondly, the

new cell state is passed through a tanh function. These two outputs are multiplied to decide what information should retain in the new hidden state. This new hidden state, together with the new cell state, form the output of a cell of the LSTM.

All cells following each other using the hidden state and cell state from the previous cell form a sequence which is able to keep the most important information throughout the whole network which makes the short term memory long again. [69]

A.6 RBFNN (Artificial Neural Network)

A neural network consists of three types of layers; the input layer, the hidden layers, and the output layer. The layers are connected and the input which travels through the network is transformed by neurons in the hidden layers to create an output. Every connection has a weight and a bias which is used to calculate the output of a hidden layer from the input of that hidden layer. An example network is shown in figure A.3. This network has one hidden layer with two neurons.

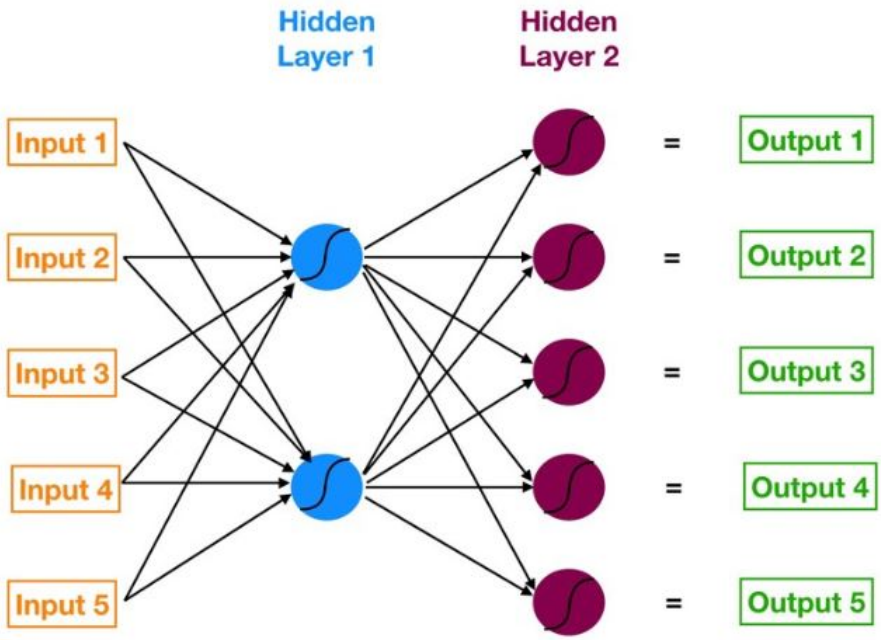


Figure A.3: An example of an ANN. Source: Figure 1 of Yiu (2019) [70]

The formula used for transforming the input is as follows:

$$Sigmoid(B_1 * X + B_0) = PredictedProbability \tag{A.11}$$

Where B_1 is the weight, X is the input, and B_0 is the bias. In this example neuron of the hidden layer, a sigmoid function is used as activation function. There are other

activation functions possible, such as the $tanh$ activation function. In every hidden layer, the input is transformed using formula A.11, however the activation function differs. All the transformations of an input by neurons in the hidden layers lead to an output. This process is called forward propagation.

The goal is to find a set of weights and biases which minimize our cost function. The cost function is a measurement of the differences between the predictions of the model and the target outcomes. There are different cost functions possible. One example is the Mean Squared Error (MSE) .

$$MSE = \sum[(Prediction - Actual)^2] * (1/num_{observations}) \tag{A.12}$$

This cost function punishes predictions with a bigger difference to the target outcome more severely.

To minimize the cost function, the derivative of the error is calculated from the output backwards to the input to be able to adjust the weights and biases. This is called backward propagation.

The forward propagation and backward propagation steps are repeated until the cost function is as minimal as possible. [70]

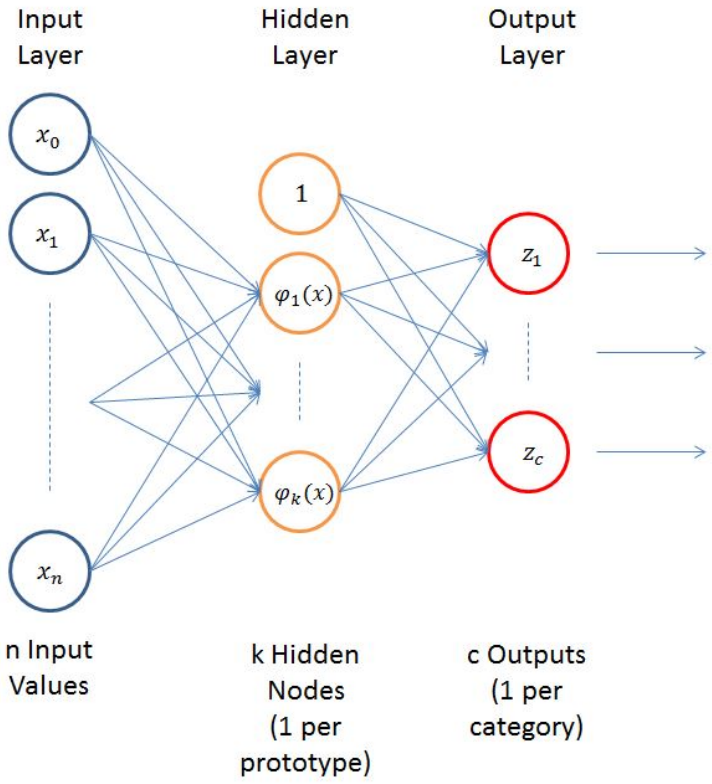


Figure A.4: The RBFNN layout. Source: figure 8 of McCormick (2013) [71]

A Radial Basis Function Neural Network (RBFNN) (see A.4) is a special form of

an ANN. It has one input layer, one hidden layer, and one output layer. The hidden layer consists of RBF neurons. These neurons store so called ”prototypes” to which the inputs are compared. The greater the Eucledian distance between the two, the lower the chance the input belongs to the same class as the prototype.

These prototypes have to be chosen in such a way that it represents a cluster of values from the same class. Therefore, K-means clustering is used to select the prototypes. The centers of these clusters will become the prototypes which is the average of all point in the cluster. The higher K, the more precise the model will be as smaller clusters are represented. However, with an increasing K, the efficiency decreases. Therefore, several values have to be tested to find the best trade-off.

The activation function of the RBF neurons (formula A.13) is based on the Gaussian.

$$\phi(x) = e^{\beta\|x-\mu\|^2} \tag{A.13}$$

The β variable controls the width of the Gaussian curve. This variable is chosen by formula A.14.

$$\beta = \frac{1}{2\sigma^2} \tag{A.14}$$

Here, the σ should be equal to the average distance between the cluster center and all points in the cluster (see formula A.15). [71]

$$\sigma = \frac{1}{m}\sum_{i=1}^{m}\|x_i - \mu\| \tag{A.15}$$

A.7 Boosted decision trees

Boosted decision trees is an ensemble model. Individual decision trees are trained sequentially and each tree learns from the mistakes made by the previous tree.

At the start of the training process, every sample has identical weights; one divided by the total number of samples. A decision tree with a depth of one is build for every feature. These decision trees are used to classify the data. The tree and feature with the highest accuracy is chosen to be the next tree in the forest. The significance of this tree is calculated using the following formula:

$$significance = \frac{1}{2}log(\frac{1 - total\_error}{total\_error}) \tag{A.16}$$

The total error is the sum of the weights of the samples which were classified incorrectly. This significance and total error is used to update the sample weight of

the samples by increasing the weights of the misclassified samples and decreasing the weights of the correctly classified samples by the following formulas:

$$new\_sample\_weight = sample\_weight * e^{significance} \tag{A.17}$$

$$new\_sample\_weight = sample\_weight * e^{-significance} \tag{A.18}$$

After that, the weights are normalized.

The new weights are used to choose a new trainingset to repeat the process of building decision trees and updating the weights. As the weights of the misclassified items is higher, the probability these will be in the new dataset multiple times is high, thus putting more emphasis on these samples.

When feeding the trees unseen data, the model classifies a new sample by each tree in the forest. The trees are divided into groups based on their output and the total significance of a group is calculated by taking the sum of the significances in this group. The output of the forest is the group with the highest total significance. [72]

A.8 Support Vector Machine

The goal of an SVM is to find a hyperplane that distinctively classifies data points.

This is done in an N-dimensional space, where N is the number of features. The to be found hyperplane should have the maximum margin. This is the distance between data points of the classes. The larger this margin is, the more confident new datapoints can be classified.

The points closest to the hyperplane are called the ”support vectors” and influence the orientation and position of the hyperplane. These support vectors are used to maximize the margin.

To decide which values are on which side of the hyperplane, the output of a linear function is examined; if this output is greater than one, it is identified with one class if it is below minus one, it is identified with the other class.

The hinge loss function together with an regularization parameter (see formula A.19) is used as the loss function for the SVM. This loss function helps in maximizing the margin. The regularization parameter balances the maximization of the margin and the loss.

$$min_w \lambda ||w||^2 + \sum_{i=1}^{n} (1 - y_i < x_i, w >)_+ \tag{A.19}$$

To find the gradients, partial derivatives with respect to the weights are taken. These gradients are used to update the weights (see formula A.20).

$$\frac{\delta}{\delta w_k}\lambda||w||^2 = 2\lambda w_k$$
$$\frac{\delta}{\delta w_k}(1 - y_i < x_i, w >)_+ = \begin{cases} 0, & if y_i < x_i, w > \geq 1 \\ -y_i x_{ik}, & else \end{cases} \tag{A.20}$$

This process leads to the maximization of the margin, finding the best hyperplane to separate the classes. [73]

A.9 Evaluation metrics

This section explains and shows example calculations of the precision, recall, F1-score, AUC-score. First, it explains the confusion matrix.

A.9.1 Confusion matrix

The example confusion matrix used for this explanation and the example calculations in this section is shown in table A.2. The confusion matrix shows how many samples in the dataset are predicted as having the label assigned to the samples, and how many samples are predicted with the wrong label. In the example confusion matrix, eighty-five samples are classified right; forty of the 'True' class, and forty-five of the 'False' class. Eight samples which have the actual label 'True' are predicted as beloning to the 'False' class. For seven attempts, it is the other way around; these have the actual label 'False', but are assigned label 'True' during prediction.

The confusion matrix is used to show the confusion between the different classes in the dataset. It shows how many attempts from one class are misclassified as another class. This is especially helpful when using a dataset with more than two classes, as it shows which of the classes are confused.

		Actual	
		True	False
Predicted	True	40	7
	False	8	45

Table A.2: Example confusion matrix

A.9.2 Precision

The formula for calculating the precision is given in equation A.21.

$$Precision = \frac{tp}{tp + fp} \tag{A.21}$$

Using the data from the example confusion matrix (A.2), the precision is calculated as follows:

$$Precision = \frac{40}{40 + 7} = \frac{40}{47} = 0.85 \tag{A.22}$$

A.9.3 Recall

To calculate the recall, formula A.23 is used.

$$Recall = \frac{tp}{tp + fn} \tag{A.23}$$

When the data from the example confusion matrix (A.2) is filled in, the recall is calculated as follows:

$$Recall = \frac{40}{40 + 8} = \frac{40}{48} = 0.83 \tag{A.24}$$

A.9.4 F1-score

The F1-score is a combination of the precision and recall, and is calculated using the following formula:

$$F1Score = \frac{2tp}{2tp + fp + fn} \tag{A.25}$$

Using the data of the example confusion matrix (A.2), the F1-score is calculated as follows:

$$F1Score = \frac{2 * 40}{2 * 40 + 7 + 8} = \frac{80}{95} = 0.84 \tag{A.26}$$

A.9.5 Precision-recall curve

The precision-recall curve shows the trade-off between recall and precision. The recall and precision are calculated at different thresholds. An example of such a curve is shown in figure A.5. Precision at 100% recall is the precision score when the recall is 1.0.

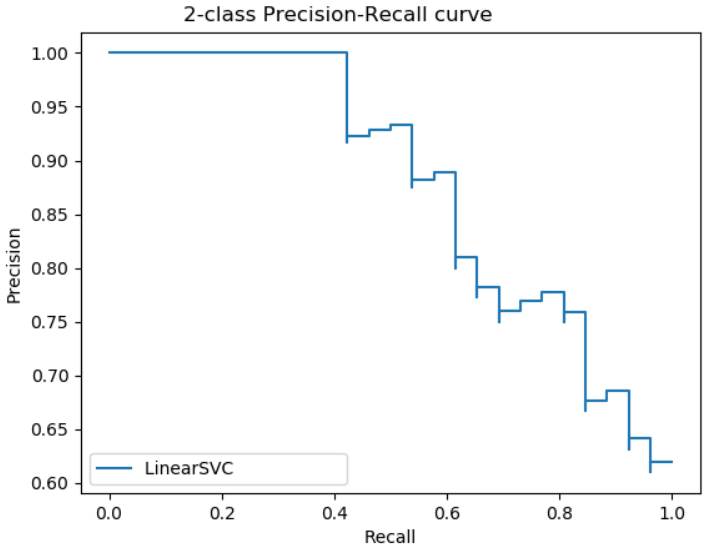


Figure A.5: An example of a precision-recall curve. Source: adapted from image 1 of Scikit Learn (2019) [74]

A.10 Statistical tests

Below are explanations and example calculations of the three statistical tests used in this research: the Shapiro-Wilk test, the independent sample T-test, and the Wilcoxon test. Also, the Q-Q plot is explained.

A.10.1 Q-Q plot

A Q-Q plot is used to compare the real quantiles of a dataset to the theoretical quantiles to visually determine if the dataset is normally distributed. This is done by dividing a curve of a normal distribution in $n+1$ segments, where n is the number of datapoints in the dataset, and finding the z-values in the z-table for the cut-off points of the segments. These z-scores are plotted against the datapoints. If this line is a straight line, the data is normally distributed. An example of a Q-Q plot where the data is approximately normally distributed is shown in figure A.6.

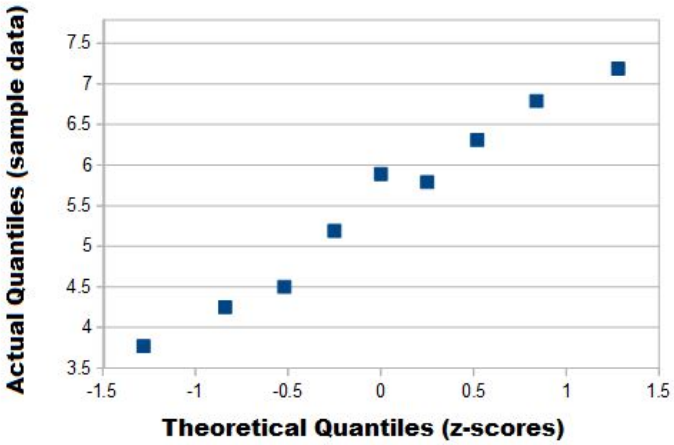


Figure A.6: An example of a Q-Q plot. Source: figure 4 of Stephanie (2019) [75]

A.10.2 Shapiro-Wilk test

To evaluate if the dependent variable is normally distributed, the Shapiro-Wilk test is performed. The formula for the Shapiro-Wilk test is as given in formula A.27.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{A.27}$$

The values (x) has to be ordered in increasing order, a_i are values to be looked up in the Shapiro-Wilk table, $\bar{x}$ is the mean of the sample. The sample is normally distributed as W is smaller or equal to c which depends on the number of entries in the sample and can also be found in the Shapiro-Wilk table.

The data in table A.3 is used as example data.

4.635	4.771	4.820	4.852	4.890	4.898	4.898	4.913	4.977	5.011
5.081	5.165	5.165	5.176	5.313	5.323	5.323	5.389	5.429	5.460
5.497	5.541	5.595	5.609	5.649	5.656	5.778	5.889	5.892	6.269

Table A.3: Example data Shapiro-Wilk test.

The calculated mean and variance are $\bar{x} = 5.295$ and $s^2 = 0.1560$. The chosen α is 0.05

H0: $F(x) = \phi(\frac{x-\mu}{\sigma})$, meaning X is normally distributed.

H1: $F(x) \neq \phi(\frac{x-\mu}{\sigma})$, meaning X is not normally distributed.

The formula used is shown in A.28, where a_i is from the table of Shapiro-Wilk.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{A.28}$$

Using the example data and the formula, $W = \frac{4.3771}{4.525} = 0.967$.

When $W \leq c$, the H0 hypothesis is not true. When looking up the value for c in the Shapiro-Wilk table when n = 30, and α = 0.05, c = 0.927, meaning $W \geq c$, so the H0 hypothesis is true.

A.10.3 Paired sample T-test

The formula used when performing an paired sample T-test is given in formula A.29.

$$T = \frac{(\sum D)/N}{\sqrt{\frac{\sum D^2 - (\frac{(\sum D)^2}{N})}{(N-1)(N)}}} \tag{A.29}$$

Where $\sum D$ is the sum of the differences, and N is the number of samples in both datasets.

The H0 hypothesis is that there are no significant differences, while the H1 hypothesis is that the differences are significant. A standard threshold of p-value 0.05 is used to determine the significance. When the difference is significant, this means there is a less than a 5% chance the difference is per accident and thus we can assume there is a real difference.

H0: the difference between the mean of the datasets is not significant. H1: the difference between the mean of the datasets is significant.

As example data we use a dataset of twenty samples with a summed difference of -73, and 11 samples in both datasets.

Filling in the formula gives us $T = -2.47$.

To find the p-value, we first need to calculate the degrees of freedom, which is $N - 1 = 10$. From the t-table we find that with an alpha level of 0.05, the t-value is 2.228. As this value is greater than our t-value at the alpha level, the p-value is less than the alpha value of 0.05. Therefore, the null hypothesis can be rejected.

A.10.4 Wilcoxon test

The Wilcoxon test is used to calculate if a difference between two groups is significant or not when the data is not normally distributed.

When performing the Wilcoxon test, the values from both samples are ordered in increasing order and the rank numbers of one sample are summed ($W = \sum iR(X_i)$). If W is bigger than the critical value chosen, there is a significant difference between the two samples.

The example data used in this example calculation is shown in table A.4.

The first step in performing the Wilcoxon test is to rank the observations. This is done in table A.10.4.

	1	2	3	4	5	6	7	8
Group 1	500	528	560	444	397	411	519	511
Group 2	410	457	501	450	407	457	385	540

Table A.4: Example data Wilcoxon test

Observation	385	397	407	410	411	444	450	451	457	500	501	511	519	528	540	560
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Belong to group 1		x			x	x				x		x	x	x		x

Table A.5: The example data ranked

The formula of the Wilcoxon test is given in formula A.30.

$$W = \sum iR(X_i) \tag{A.30}$$

H0: $f_X(x) = f_Y(x)$

H1: $f_X(x) = f_Y(x-a)$, such that $a > 0$, and $\alpha_0 = 0.10$.

Using the ranked example data (table A.10.4), W is calculated as follows:

$$W = \sum_{i=1}^{8} R(X_i) = 2 + 5 + 6 + 10 + 12 + 13 + 14 + 16 = 78 \tag{A.31}$$

As the number of observations in each group is bigger than five, we are allowed to say that W is approximately normally distributed where:

$\mu = E(W) = \frac{1}{2}n_1(N+1) = \frac{1}{2} * 8 * (16+1) = 68$, and

$\sigma^2 = \frac{1}{12}n_1n_2(N+1) = \frac{1}{12} * 8 * 8 * (16+1) \approx 90.67$

H0 is not true if $P(W \geq 78|H_0) \leq \alpha_0$:

$P(W \geq 78|H_0) = P(W \geq 77.5|H_0) \approx P(Z \geq \frac{77.5-68}{\sqrt{90.67}}) \approx 1 - \phi(1.00) = 15.87\%$

15.87% $>$ 10% = α_0, so the difference between the two groups is not significant.

Appendix B

Features

B.1 Unfiltered features

Feature	Explanation
abs_energy(x)	Returns the absolute energy of the time series which is the sum over the squared values.
absolute_sum_of_changes(x)	Returns the sum over the absolute value of consecutive changes in the series x.
agg_autocorrelation(x, param)	Calculates the value of an aggregation function over the autocorrelation for the different lags.
agg_linear_trend(x, param)	Calculates a linear least-squares regression for values of the time series that were aggregated over chunks versus the sequence from 0 up to the number of chunks minus one.
approximate_entropy(x, m, r)	Implements a vectorized Approximate entropy algorithm.
ar_coefficient(x, param)	This feature calculator fits the unconditional maximum likelihood of an autoregressive AR(k) process.
augmented_dickey_fuller(x, param)	The Augmented Dickey-Fuller test is a hypothesis test which checks whether a unit root is present in a time series sample.
autocorrelation(x, lag)	Calculates the autocorrelation of the specified lag.
benford_correlation(x)	Returns the correlation from first digit distribution when compared to the Newcomb-Benfords Law distribution.
binned_entropy(x, max_bins)	First bins the values of x into max_bins equidistant bins. The calculates the sum over $p_k \log(p_k) * 1_{(p_k>0)}$.
c3(x, lag)	A measurement of the non-linearity in the time-series.
change_quantiles(x, ql, qh, isabs, f_agg)	Calculates the average, absolute value of consecutive changes of the series x inside a fixed corridor given by the quantiles ql and qh of the distribution of x.

Table B.1 continued from previous page

Feature	Explanation
cid_ce(x, normalize)	This function calculator is an estimate for a time series complexity
count_above(x, t)	Returns the percentage of values in x that are higher than t.
count_above_mean(x)	Returns the number of values in x that are higher than the mean of x.
count_below(x, t)	Returns the percentage of values in x that are lower than t.
count_below_mean(x)	Returns the number of values in x that are lower than the mean of x.
cwt_coefficients(x, param)	Calculates a Continuous wavelet transform for the Ricker wavelet, also known as the Mexican hat wavelet.
energy_ratio_by_chunks(x, param)	Calculates the sum of squares of chunk i out of N chunks expressed as a ratio with the sum of squares over the whole series.
fft_aggregated(x, param)	Returns the spectral centroid (mean), variance, skew, and kurtosis of the absolute fourier transform spectrum.
fft_coefficient(x, param)	Calculates the fourier coefficients of the one-dimensional discrete Fourier Transform for real input.
first_location_of_maximum(x)	Returns the first location of the maximum value of x.
first_location_of_minimum(x)	Returns the first location of the minimal value of x.
fourier_entropy(x, bins)	Calculate the binned entropy of the power spectral density of the time series (using the welch method).
friedrich_coefficients(x, param)	Coefficients of polynomial $h(x)$, which has been fitted to the deterministic dynamics of Langevin model.

Table B.1 continued from previous page

Feature	Explanation
has_duplicate(x)	Checks if any value in x occurs more than once.
has_duplicate_max(x)	Checks if the maximum value of x is observed more than once.
has_duplicate_min(x)	Checks if the minimal value of x is observed more than once.
index_mass_quantile(x, param)	Calculate the relative index i where q% of the mass of the time series x lie left of i.
kurtosis(x)	Returns the kurtosis of x (calculated with the adjusted Fisher-Pearson standardized moment coefficient G2).
large_standard_deviation(x, r)	Boolean variable denoting if the standard dev of x is higher than r times the range = difference between max and min of x.
last_location_of_maximum(x)	Returns the relative last location of the maximum value of x.
last_location_of_minimum(x)	Returns the last location of the minimal value of x.
lempel_ziv_complexity(x, bins)	Calculate a complexity estimate based on the Lempel-Ziv compression algorithm.
length(x)	Returns the length of x.
linear_trend(x, param)	Calculate a linear least-squares regression for the values of the time series versus the sequence from 0 to length of the time series minus one.
linear_trend_timewise(x, param)	Calculate a linear least-squares regression for the values of the time series versus the sequence from 0 to length of the time series minus one.
longest_strike_above_mean(x)	Returns the length of the longest consecutive subsequence in x that is bigger than the mean of x.

Table B.1 continued from previous page

Feature	Explanation
longest_strike_below_mean(x)	Returns the length of the longest consecutive subsequence in x that is smaller than the mean of x.
max_langevin_fixed_point(x, r, m)	Largest fixed point of dynamics :math:argmax_x {h(x)=0}‘ estimated from polynomial $h(x)$, which has been fitted to the deterministic dynamics of the Langevin model.
maximum(x)	Calculates the highest value of the time series x.
mean(x)	Returns the mean of x.
mean_abs_change(x)	Returns the mean over the absolute differences between subsequent time series values.
mean_change(x)	Returns the mean over the differences between subsequent time series values.
mean_second_derivative_central(x)	Returns the mean value of a central approximation of the second derivative.
median(x)	Returns the median of x.
minimum(x)	Calculates the lowest value of the time series x.
number_crossing_m(x, m)	Calculates the number of crossings of x on m.
number_cwt_peaks(x, n)	Calculates the number of peaks that occur at enough width scales and with sufficiently high Signal-to-Noise-Ratio (SNR).
number_peaks(x, n)	Calculates the number of peaks of at least support n in the time series x.
partial_autocorrelation(x, param)	Calculates the value of the partial autocorrelation function at the given lag.

Table B.1 continued from previous page

Feature	Explanation
percentage_of_reoccurring_datapoints_to_all_datapoints(x)	Returns the percentage of non-unique values.
percentage_of_reoccurring_values_to_all_values(x)	Returns the ratio of values that are present in the time series more than once.
permutation_entropy(x, tau, dimension)	Calculate the permutation entropy.
quantile(x, q)	Calculates the q quantile of x.
range_count(x, min, max)	Count observed values within the interval [min, max).
ratio_beyond_r_sigma(x, r)	Ratio of values that are more than r*std(x) (so r sigma) away from the mean of x.
ratio_value_number_to_time_series_length(x)	Returns a factor which is 1 if all values in the time series occur only once, and below one if this is not the case.
sample_entropy(x)	Calculates and returns sample entropy of x.
set_property(key, value)	This method returns a decorator that sets the property key of the function to value.
skewness(x)	Returns the sample skewness of x (calculated with the adjusted Fisher-Pearson standardized moment coefficient G1).
spkt_welch_density(x, param)	This feature calculator estimates the cross power spectral density of the time series x at different frequencies.
standard_deviation(x)	Returns the standard deviation of x.
sum_of_reoccurring_data_points(x)	Returns the sum of all data points, that are present in the time series more than once.
sum_of_reoccurring_values(x)	Returns the sum of all values, that are present in the time series more than once.

Table B.1 continued from previous page

Feature	Explanation
sum_values(x)	Calculates the sum over the time series values.
symmetry_looking(x, param)	Boolean variable denoting if the distribution of x looks symmetric.
time_reversal_asymmetry_statistic(x, lag)	This function calculates the value of $E[L^2(X)^2 * L(X) - L(X) * X^2]$, where E is the mean, and L is the lag operator.
value_count(x, value)	Count occurrences of value in time series x.
variance(x)	Returns the variance of x.
variance_larger_than_standard_deviation(x)	Boolean variable denoting if the variance of x is greater than its standard deviation.
variation_coefficient(x)	Returns the variation coefficient (standard error / mean, give relative value of variation around mean) of x.

Table B.1: The unfiltered featureset, adapted from Christ et al. (2020) [76].

B.2 The filtered features for the binary labelset

Feature	Explanation
0__change_quantiles__f_agg_"var"__isabs_False__qh_1.0__ql_0.8	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.8 to 1.0, using the variability as aggregation function, and the non-absolute values.
0__change_quantiles__f_agg_"mean"__isabs_True__qh_1.0__ql_0.8	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.8 to 1.0, using the mean as aggregation function, and the absolute values.
0__change_quantiles__f_agg_"var"__isabs_True__qh_1.0__ql_0.8	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.8 to 1.0, using the variability as aggregation function, and the absolute values.
0__ar_coefficient__k_10__coeff_1	The fitting of an unconditional maximum likelihood of an autoregressive process, containing a maximum lag of ten, and a φ of one
0__ar_coefficient__k_10__coeff_4	The fitting of an unconditional maximum likelihood of an autoregressive process, containing a maximum lag of ten, and a φ of four
0__ar_coefficient__k_10__coeff_2	The fitting of an unconditional maximum likelihood of an autoregressive process, containing a maximum lag of ten, and a φ of two
0__ar_coefficient__k_10__coeff_3	The fitting of an unconditional maximum likelihood of an autoregressive process, containing a maximum lag of ten, and a φ of three

Table B.2 continued from previous page

Feature	Explanation
0__change_quantiles__f_agg_"var"__isabs_False__qh_0.2__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.2, using the variability as aggregation function, and the non-absolute values.
0__spkt_welch_density__coeff_5	The cross power spectral density with coefficient five.
0__agg_linear_trend__f_agg_"min"__chunk_len_5__attr_"stderr"	A linear least squares regression with aggregation function minimum, a chunk length of five, and the extracted attribute is the standard error.
0__agg_linear_trend__f_agg_"min"__chunk_len_10__attr_"stderr"	A linear least squares regression with aggregation function minimum, a chunk length of ten, and the extracted attribute is the standard error.
0__change_quantiles__f_agg_"mean"__isabs_True__qh_0.2__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.2, using the mean as aggregation function, and the absolute values.
0__change_quantiles__f_agg_"mean"__isabs_True__qh_0.6__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.6, using the mean as aggregation function, and the absolute values.
0__agg_linear_trend__f_agg_"min"__chunk_len_50__attr_"intercept"	A linear least squares regression with aggregation function minimum, a chunk length of fifty, and the extracted attribute is the intercept.

Table B.2 continued from previous page

Feature	Explanation
0__agg_linear_trend__f_agg_"min"__chunk_len_50__attr_"slope"	A linear least squares regression with aggregation function minimum, a chunk length of fifty, and the extracted attribute is the slope.
0__change_quantiles__f_agg_"mean"__isabs_True__qh_0.8__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.8, using the mean as aggregation function, and the absolute values.
0__change_quantiles__f_agg_"var"__isabs_True__qh_0.2__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.8 to 1.0, using the variability as aggregation function, and the non-absolute values.
0__range_count__max_1__min_-1	The number of values between -1 and 1.
0__agg_linear_trend__f_agg_"min"__chunk_len_50__attr_"stderr"	A linear least squares regression with aggregation function minimum, a chunk length of fifty, and the extracted attribute is the standard error.
0__change_quantiles__f_agg_"mean"__isabs_True__qh_0.4__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.4, using the mean as aggregation function, and the absolute values.
0__change_quantiles__f_agg_"var"__isabs_False__qh_0.6__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.6, using the variability as aggregation function, and the non-absolute values.

Table B.2 continued from previous page

Feature	Explanation
0__change_quantiles__f_agg_"var"__isabs_False__qh_0.4__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.4, using the variability as aggregation function, and the non-absolute values.
0__minimum	The lowest value of the timeseries.
0__change_quantiles__f_agg_"var"__isabs_True__qh_0.6__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.6, using the variability as aggregation function, and the absolute values.
0__change_quantiles__f_agg_"var"__isabs_True__qh_0.4__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.4, using the variability as aggregation function, and the absolute values.
0__quantile__q_0.1	The 0.1 quantile.
0__number_crossing_m__m_-1	The number of crossing of the -1 line.
0__change_quantiles__f_agg_"var"__isabs_False__qh_0.8__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.8, using the variability as aggregation function, and the non-absolute values.
0__change_quantiles__f_agg_"var"__isabs_True__qh_1.0__ql_0.6	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.6 to 1.0, using the variability as aggregation function, and the absolute values.

Table B.2 continued from previous page

Feature	Explanation
0__change_quantiles__f_agg_"var"__isabs_True__qh_0.8__ql_0.0	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.0 to 0.8, using the variability as aggregation function, and the absolute values.
0__change_quantiles__f_agg_"var"__isabs_False__qh_1.0__ql_0.6	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.6 to 1.0, using the variability as aggregation function, and the non-absolute values.
0__change_quantiles__f_agg_"mean"__isabs_False__qh_1.0__ql_0.8	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.8 to 1.0, using the mean as aggregation function, and the non-absolute values.
0__spkt_welch_density__coeff_2	The cross power spectral density with coefficient two
0__agg_linear_trend__f_agg_"min"__chunk_len_10__attr_"intercept"	A linear least squares regression with aggregation function minimum, a chunk length of ten, and the extracted attribute is the intercept.
0__change_quantiles__f_agg_"mean"__isabs_True__qh_1.0__ql_0.4	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.4 to 1.0, using the mean as aggregation function, and the absolute values.

Table B.2 continued from previous page

Feature	Explanation
0__change_quantiles__f_agg_"var"__isabs_False__qh_1.0__ql_0.4	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.4 to 1.0, using the variability as aggregation function, and the non-absolute values.
0__change_quantiles__f_agg_"var"__isabs_True__qh_1.0__ql_0.4	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.4 to 1.0, using the variability as aggregation function, and the absolute values.
0__ar_coefficient__k_10__coeff_0	The fitting of an unconditional maximum likelihood of an autoregressive process, containing a maximum lag of ten, and a φ of zero.
0__fft_coefficient__coeff_2__attr_"imag"	The fourier coefficients of the 1D discrete Fourier Transform with coefficient two, and returning the imaginary part.
0__spkt_welch_density__coeff_8	The cross power spectral density with coefficient eight.
0__change_quantiles__f_agg_"mean"__isabs_True__qh_1.0__ql_0.6	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.6 to 1.0, using the mean as aggregation function, and the absolute values.
0__agg_linear_trend__f_agg_"min"__chunk_len_10__attr_"slope"	A linear least squares regression with aggregation function minimum, a chunk length of ten, and the extracted attribute is the slope.

Table B.2 continued from previous page

Feature	Explanation
0__agg_linear_trend__f_agg_"min"__chunk_len_5__attr_"intercept"	A linear least squares regression with aggregation function minimum, a chunk length of five, and the extracted attribute is the intercept.
0__agg_linear_trend__f_agg_"min"__chunk_len_5__attr_"slope"	A linear least squares regression with aggregation function minimum, a chunk length of five, and the extracted attribute is the slope.
0__percentage_of_reoccurring_datapoints_to_all_datapoints	The percentage of non-unique datapoints
0__change_quantiles__f_agg_"var"__isabs_True__qh_0.8__ql_0.2	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.2 to 0.8, using the variability as aggregation function, and the absolute values.
0__change_quantiles__f_agg_"var"__isabs_False__qh_0.8__ql_0.2	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.2 to 0.8, using the variability as aggregation function, and the non-absolute values.
0__c3__lag_3	The non linearity, with a lag of three.
0__fft_coefficient__coeff_3__attr_"abs"	The fourier coefficients of the 1D discrete Fourier Transform with coefficient three, and returning the absolute value.
0__count_below__t_0	The percentage of values lower than zero.
0__linear_trend__attr_"pvalue"	A linear least-squares regression, returning the p-value
0__quantile__q_0.9	The 0.9 quantile.
0__c3__lag_2	The non linearity, with a lag of two.

Table B.2 continued from previous page

Feature	Explanation
0__fft_coefficient__coeff_21__attr_"angle"	The fourier coefficients of the 1D discrete Fourier Transform with coefficient twenty-one, and returning the angle.
0__change_quantiles__f_agg_"mean"__isabs_False__qh_0.8__ql_0.4	The average value of consecutive changes of the data inside the quantile. This value is calculated for quantile 0.4 to 0.8, using the mean as aggregation function, and the absolute values.
0__fft_coefficient__coeff_21__attr_"imag"	The fourier coefficients of the 1D discrete Fourier Transform with coefficient twenty-one, and returning the imaginary part.

Table B.2: The features in the filtered featureset for the best performing binary model, adapted from Christ et al. (2020) [76].

Appendix C

Performance of the models

C.1 Hyperparameter tuning

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.639	0.525	0.577	0.539
	Boosted decision trees	0.726	0.653	0.688	0.532
	SVM	0.914	0.720	0.806	0.605
Filtered features	RBFNN	0.629	0.559	0.592	0.500
	Boosted decision trees	0.699	0.610	0.652	0.541
	SVM	0.763	0.737	0.750	0.573
Filtered features + spirometry parameters	RBFNN	0.602	0.500	0.546	0.500
	Boosted decision trees	0.623	0.559	0.589	0.504
	SVM	0.733	0.720	0.726	0.557
Unfiltered features	RBFNN	0.592	0.381	0.464	0.500
	Boosted decision trees	0.566	0.475	0.516	0.509
	SVM	0.632	0.508	0.563	0.518
Unfiltered features + spirometry parameters	RBFNN	0.618	0.466	0.531	0.500
	Boosted decision trees	0.525	0.449	0.484	0.504
	SVM	0.634	0.500	0.559	0.511
Smoothed data	LSTM	0.384	0.449	0.414	0.500

Table C.1: The performance of the models after hyperparameter tuning, using labelset 'binary'.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.380	0.203	0.258	0.117
	Boosted decision trees	0.410	0.229	0.293	0.113
	SVM	0.393	0.356	0.373	0.125
Filtered features	RBFNN	0.198	0.178	0.177	0.111
	Boosted decision trees	0.301	0.136	0.179	0.111
	SVM	0.362	0.263	0.304	0.111
Filtered features + spirometry parameters	RBFNN	0.173	0.127	0.146	0.111
	Boosted decision trees	0.242	0.169	0.199	0.111
	SVM	0.372	0.229	0.282	0.115
Unfiltered features	RBFNN	0.044	0.017	0.025	0.113
	Boosted decision trees	0.034	0.017	0.023	0.111
	SVM	0.165	0.051	0.077	0.111
Unfiltered features + spirometry parameters	RBFNN	0.141	0.042	0.065	0.111
	Boosted decision trees	0.119	0.059	0.077	0.111
	SVM	0.131	0.059	0.081	0.111
Smoothed data	LSTM	0.097	0.102	0.097	0.112

Table C.2: The performance of the models after hyperparameter tuning, using labelset 'combined'.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.277	0.178	0.193	0.065
	Boosted decision trees	0.191	0.119	0.146	0.063
	SVM	0.270	0.305	0.286	0.063
Filtered features	RBFNN	0.184	0.144	0.150	0.063
	Boosted decision trees	0.120	0.059	0.078	0.063
	SVM	0.233	0.195	0.212	0.063
Filtered features + spirometry parameters	RBFNN	0.160	0.161	0.148	0.063
	Boosted decision trees	0.163	0.144	0.150	0.063
	SVM	0.211	0.195	0.202	0.063
Unfiltered features	RBFNN	0.086	0.025	0.039	0.063
	Boosted decision trees	0.048	0.025	0.033	0.063
	SVM	0.000	0.000	0.000	0.063
Unfiltered features + spirometry parameters	RBFNN	0.064	0.017	0.027	0.063
	Boosted decision trees	0.034	0.025	0.028	0.063
	SVM	0.000	0.000	0.000	0.063
Smoothed data	LSTM	0.045	0.127	0.066	0.063

Table C.3: The performance of the models after hyperparameter tuning, using labelset 'all'.

C.2 Balancing

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.744	0.568	0.644	0.524
	Boosted decision trees	0.776	0.644	0.704	0.524
	SVM	0.874	0.763	0.814	0.567
Filtered features	RBFNN	0.682	0.636	0.658	0.534
	Boosted decision trees	0.675	0.653	0.664	0.534
	SVM	0.760	0.780	0.770	0.573
Filtered features + spirometry parameters	RBFNN	0.615	0.610	0.613	0.518
	Boosted decision trees	0.620	0.636	0.628	0.513
	SVM	0.738	0.763	0.750	0.562
Unfiltered features	RBFNN	0.608	0.525	0.564	0.500
	Boosted decision trees	0.549	0.568	0.558	0.500
	SVM	0.585	0.525	0.554	0.524
Unfiltered features + spirometry parameters	RBFNN	0.570	0.551	0.560	0.500
	Boosted decision trees	0.551	0.551	0.551	0.500
	SVM	0.630	0.534	0.578	0.500
Smoothed data	LSTM	0.371	0.441	0.403	0.509

Table C.4: The performance of the models, after applying balancing technique SMOTE, using labelset 'binary'.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.702	0.559	0.623	0.541
	Boosted decision trees	0.771	0.686	0.726	0.596
	SVM	0.849	0.763	0.804	0.615
Filtered features	RBFNN	0.596	0.576	0.586	0.529
	Boosted decision trees	0.652	0.619	0.635	0.504
	SVM	0.857	0.864	0.861	0.678
Filtered features + spirometry parameters	RBFNN	0.661	0.644	0.652	0.518
	Boosted decision trees	0.649	0.610	0.629	0.502
	SVM	0.741	0.729	0.735	0.599
Unfiltered features	RBFNN	0.566	0.542	0.554	0.502
	Boosted decision trees	0.525	0.542	0.533	0.502
	SVM	0.583	0.508	0.543	0.522
Unfiltered features + spirometry parameters	RBFNN	0.625	0.593	0.609	0.502
	Boosted decision trees	0.565	0.517	0.540	0.500
	SVM	0.562	0.619	0.589	0.527
Smoothed data	LSTM	0.440	0.559	0.493	0.500

Table C.5: The performance of the models, after applying balancing technique ROS, using labelset 'binary'.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.342	0.246	0.272	0.120
	Boosted decision trees	0.363	0.263	0.267	0.111
	SVM	0.380	0.246	0.287	0.122
Filtered features	RBFNN	0.205	0.195	0.194	0.111
	Boosted decision trees	0.275	0.203	0.220	0.113
	SVM	0.326	0.271	0.285	0.121
Filtered features + spirometry parameters	RBFNN	0.234	0.153	0.179	0.111
	Boosted decision trees	0.185	0.186	0.175	0.118
	SVM	0.251	0.203	0.222	0.115
Unfiltered features	RBFNN	0.090	0.068	0.073	0.112
	Boosted decision trees	0.070	0.119	0.088	0.114
	SVM	0.163	0.186	0.173	0.112
Unfiltered features + spirometry parameters	RBFNN	0.096	0.076	0.074	0.111
	Boosted decision trees	0.106	0.186	0.135	0.111
	SVM	0.131	0.127	0.128	0.112
Smoothed data	LSTM	0.085	0.127	0.096	0.111

Table C.6: The performance of the models, after applying balancing technique SMOTE, using labelset 'combined'.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.385	0.288	0.304	0.122
	Boosted decision trees	0.292	0.305	0.284	0.120
	SVM	0.584	0.525	0.547	0.134
Filtered features	RBFNN	0.267	0.246	0.252	0.111
	Boosted decision trees	0.275	0.237	0.246	0.113
	SVM	0.172	0.144	0.153	0.113
Filtered features + spirometry parameters	RBFNN	0.195	0.161	0.166	0.111
	Boosted decision trees	0.234	0.305	0.256	0.115
	SVM	0.214	0.153	0.164	0.113
Unfiltered features	RBFNN	0.108	0.076	0.081	0.112
	Boosted decision trees	0.116	0.169	0.135	0.113
	SVM	0.163	0.153	0.157	0.112
Unfiltered features + spirometry parameters	RBFNN	0.144	0.119	0.105	0.114
	Boosted decision trees	0.102	0.153	0.120	0.112
	SVM	0.129	0.127	0.124	0.112
Smoothed data	LSTM	0.120	0.153	0.128	0.113

Table C.7: The performance of the models, after applying balancing technique ROS, using labelset 'combined'.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.226	0.220	0.199	0.064
	Boosted decision trees	0.200	0.237	0.209	0.063
	SVM	0.324	0.297	0.305	0.063
Filtered features	RBFNN	0.122	0.161	0.132	0.063
	Boosted decision trees	0.196	0.263	0.208	0.063
	SVM	0.225	0.237	0.229	0.063
Filtered features + spirometry parameters	RBFNN	0.106	0.127	0.111	0.063
	Boosted decision trees	0.179	0.220	0.188	0.063
	SVM	0.188	0.178	0.180	0.063
Unfiltered features	RBFNN	0.005	0.008	0.006	0.064
	Boosted decision trees	0.045	0.085	0.058	0.063
	SVM	0.052	0.051	0.052	0.063
Unfiltered features + spirometry parameters	RBFNN	0.021	0.042	0.026	0.063
	Boosted decision trees	0.029	0.051	0.035	0.063
	SVM	0.049	0.042	0.043	0.063
Smoothed data	LSTM	0.027	0.059	0.033	0.063

Table C.8: The performance of the models, after applying balancing technique SMOTE, using labelset 'all'.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.209	0.220	0.196	0.065
	Boosted decision trees	0.190	0.212	0.195	0.063
	SVM	0.340	0.322	0.321	0.063
Filtered features	RBFNN	0.137	0.144	0.123	0.063
	Boosted decision trees	0.141	0.178	0.152	0.063
	SVM	0.249	0.254	0.249	0.063
Filtered features + spirometry parameters	RBFNN	0.125	0.093	0.090	0.063
	Boosted decision trees	0.209	0.246	0.208	0.063
	SVM	0.189	0.186	0.183	0.063
Unfiltered features	RBFNN	0.015	0.017	0.016	0.064
	Boosted decision trees	0.069	0.127	0.087	0.063
	SVM	0.033	0.042	0.037	0.063
Unfiltered features + spirometry parameters	RBFNN	0.159	0.034	0.033	0.063
	Boosted decision trees	0.131	0.169	0.143	0.063
	SVM	0.023	0.025	0.024	0.063
Smoothed data	LSTM	0.076	0.085	0.074	0.063

Table C.9: The performance of the models, after applying balancing technique ROS, using labelset 'all'.

C.3 Proposed decision tree

C.3.1 Labelset: combined without the 0 errorclass

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.105	0.088	0.095	0.128
	Boosted decision trees	0.197	0.138	0.158	0.131
	SVM	0.159	0.075	0.102	0.125
Filtered features	RBFNN	0.312	0.213	0.213	0.130
	Boosted decision trees	0.171	0.163	0.164	0.126
	SVM	0.255	0.150	0.189	0.126
Filtered features + spirometry parameters	RBFNN	0.131	0.150	0.140	0.127
	Boosted decision trees	0.197	0.138	0.158	0.131
	SVM	0.159	0.075	0.102	0.125
Unfiltered features	RBFNN	0.156	0.163	0.140	0.128
	Boosted decision trees	0.156	0.163	0.151	0.128
	SVM	0.188	0.188	0.177	0.125
Unfiltered features + spirometry parameters	RBFNN	0.119	0.200	0.149	0.127
	Boosted decision trees	0.173	0.150	0.150	0.125
	SVM	0.131	0.150	0.138	0.125
Smoothed data	LSTM	0.089	0.063	0.073	0.126

Table C.10: The performance of the models, using the imbalanced dataset and the combined labelset without the zero errorclass.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.209	0.175	0.159	0.130
	Boosted decision trees	0.123	0.100	0.080	0.126
	SVM	0.177	0.113	0.127	0.125
Filtered features	RBFNN	0.342	0.100	0.112	0.127
	Boosted decision trees	0.092	0.113	0.078	0.125
	SVM	0.274	0.188	0.216	0.125
Filtered features + spirometry parameters	RBFNN	0.194	0.113	0.131	0.127
	Boosted decision trees	0.117	0.100	0.086	0.127
	SVM	0.195	0.150	0.156	0.126
Unfiltered features	RBFNN	0.148	0.088	0.099	0.125
	Boosted decision trees	0.168	0.188	0.176	0.128
	SVM	0.116	0.138	0.126	0.126
Unfiltered features + spirometry parameters	RBFNN	0.168	0.150	0.158	0.125
	Boosted decision trees	0.142	0.163	0.151	0.129
	SVM	0.118	0.113	0.115	0.126
Smoothed data	LSTM	0.141	0.063	0.118	0.126

Table C.11: The performance of the models, after applying SMOTE, using the combined labelset without the zero errorclass.

Featureset	**Model**	**Precision**	**Recall**	**F1-score**	**Precision at 100% recall**
Spirometry parameters	RBFNN	0.131	0.088	0.076	0.130
	Boosted decision trees	0.168	0.138	0.148	0.135
	SVM	0.192	0.150	0.156	0.127
Filtered features	**RBFNN**	**0.176**	**0.225**	**0.198**	**0.131**
	Boosted decision trees	**0.343**	**0.338**	**0.332**	**0.127**
	SVM	0.309	0.163	0.185	0.125
Filtered features + spirometry parameters	RBFNN	0.189	0.200	0.171	0.127
	Boosted decision trees	0.132	0.138	0.135	0.127
	SVM	**0.195**	**0.188**	**0.177**	**0.128**
Unfiltered features	RBFNN	0.237	0.138	0.144	0.125
	Boosted decision trees	0.177	0.225	0.198	0.125
	SVM	0.170	0.150	0.159	0.127
Unfiltered features + spirometry parameters	RBFNN	0.189	0.175	0.172	0.126
	Boosted decision trees	0.230	0.163	0.190	0.126
	SVM	0.182	0.175	0.178	0.126
Smoothed data	LSTM	0.138	0.075	0.059	0.126

Table C.12: The performance of the models, after applying ROS, using the combined labelset without the zero errorclass.

C.3.2 Labelset: 10 to 20, and 66

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.033	0.050	0.040	0.146
	Boosted decision trees	0.020	0.050	0.029	0.125
	SVM	0	0	0	0.125
Filtered features	RBFNN	0.100	0.050	0.067	0.132
	Boosted decision trees	0	0	0	0.125
	SVM	0	0	0	0.125
Filtered features + spirometry parameters	RBFNN	0.050	0.050	0.050	0.130
	Boosted decision trees	0.020	0.050	0.029	0.125
	SVM	0	0	0	0.125
Unfiltered features	RBFNN	0.050	0.050	0.050	0.127
	Boosted decision trees	0	0	0	0.125
	SVM	0	0	0	0.125
Unfiltered features + spirometry parameters	RBFNN	0.038	0.050	0.043	0.131
	Boosted decision trees	0	0	0	0.125
	SVM	0	0	0	0.125
Smoothed data	LSTM	0.150	0.050	0.075	0.125

Table C.13: The performance of the models trained and evaluated on the dataset consisting of the attempts with labels between 10 and 20, and 66.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	RBFNN	0.233	0.100	0.120	0.127
	Boosted decision trees	0.110	0.250	0.152	0.125
	SVM	0.022	0.050	0.031	0.125
Filtered features	RBFNN	0.142	0.100	0.117	0.136
	Boosted decision trees	0	0	0	0.125
	SVM	**0.120**	**0.150**	**0.133**	**0.125**
Filtered features + spirometry parameters	RBFNN	0.021	0.050	0.030	0.126
	Boosted decision trees	0.040	0.100	0.057	0.125
	SVM	0.080	0.100	0.089	0.125
Unfiltered features	RBFNN	0	0	0	0.127
	Boosted decision trees	0.076	0.250	0.106	0.125
	SVM	0.200	0.050	0.080	0.125
Unfiltered features + spirometry parameters	RBFNN	0.038	0.050	0.043	0.135
	Boosted decision trees	0.048	0.100	0.064	0.125
	SVM	0.129	0.100	0.103	0.125
Smoothed data	LSTM	0.029	0.050	0.036	0.125

Table C.14: The performance of the models trained and evaluated on the dataset consisting of the attempts with labels between 10 and 20, and 66, after applying balancing technique SMOTE.

Featureset	Model	Precision	Recall	F1-score	Precision at 100% recall
Spirometry parameters	**RBFNN**	**0.168**	**0.200**	**0.182**	**0.142**
	Boosted decision trees	0.082	0.250	0.122	0.125
	SVM	0	0	0	0.125
Filtered features	RBFNN	0.050	0.050	0.050	0.133
	Boosted decision trees	**0.350**	**0.300**	**0.305**	**0.125**
	SVM	0.100	0.050	0.067	0.125
Filtered features + spirometry parameters	RBFNN	0.357	0.200	0.219	0.127
	Boosted decision trees	0.015	0.050	0.024	0.125
	SVM	0.100	0.100	0.100	0.125
Unfiltered features	RBFNN	0.055	0.100	0.071	0.129
	Boosted decision trees	0.065	0.100	0.073	0.125
	SVM	0.100	0.100	0.097	0.125
Unfiltered features + spirometry parameters	RBFNN	0	0	0	0.127
	Boosted decision trees	0.207	0.300	0.205	0.125
	SVM	0.095	0.100	0.094	0.125
Smoothed data	**LSTM**	**0.082**	**0.150**	**0.106**	**0.125**

Table C.15: The performance of the models trained and evaluated on the dataset consisting of the attempts with labels between 10 and 20, and 66, after applying balancing techinque ROS.

C.3.3 Stacking

The models in bold in tables C.10 to C.15 were stacked. Although the precision at 100% recall was higher for another model of the boosted decision trees using the combined labelset without the zero errorclass, and the RBFNN using the labelset including the labels eight to twenty, and sixty-six, the difference in recall was always bigger and thus these models were used in stacking. Additionally, the boosted decision trees using the labelset including the labels eight to twenty, and sixty-six, and the filtered featureset performed equally compared to the model using the unfiltered featureset in combination with the spirometry parameters, based on recall and precision at 100% recall. In this case, the precision score was used to determine which model to use.

Tables C.16 to C.19 show the performance of the models on the train and testset, and the correlation between the single models of the stacked models. The performance on the testset is lower than the performance of the single models. The performance on the trainingset is very high for both models. Table C.17 shows that the correlation between the single models of the stacked model for the second stage of the proposed decision tree is low. The correlation between the single models of the stacked model for the third stage (table C.18) are higher, except for the correlation between the LSTM and the other models.

Labelset	**Precision**	**Recall**	**F1-score**	**Precision at 100% recall**
Combined, without the zero errorclass (stage 2)	1	1	1	1
Attempts with labels 8 to 20, and 66 (stage 3)	0.908	0.918	0.909	0.788

Table C.16: The performance of the stacked models for stage two and three of the proposed decision tree, evaluated on the trainingset.

	RBFNN	**Boosted decision trees**
Boosted decision trees	0.200	
SVM	0.358	0.305

Table C.17: The correlation between the different single models used in stacking for the second stage of the proposed decision tree.

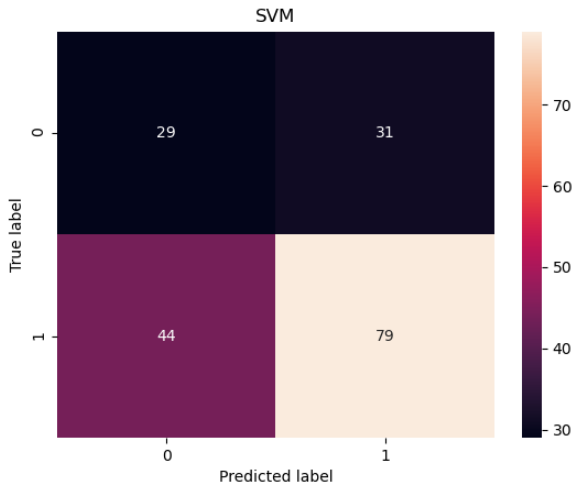
	RBFNN	**Boosted decision trees**	**SVM**
Boosted decision trees	0.503		
SVM	0.746	0.585	
LSTM	0.046	0.142	0.038

Table C.18: The correlation between the different models used in stacking for the third stage of the proposed decision tree.

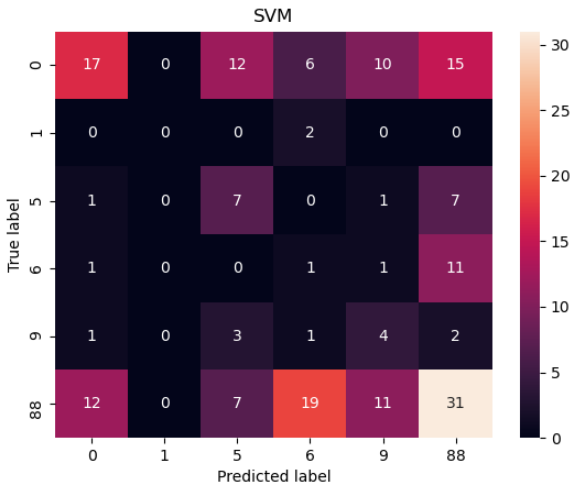
	Precision	**Recall**	**F1-score**	**Precision at 100% recall**
Combined, without the zero errorclass (stage 2)	0.252	0.200	0.221	0.125
Attempts with labels 8 to 20, and 66 (stage 3)	0.101	0.150	0.119	0.125

Table C.19: The performance of the stacked models for stage two and three of the proposed decision tree.

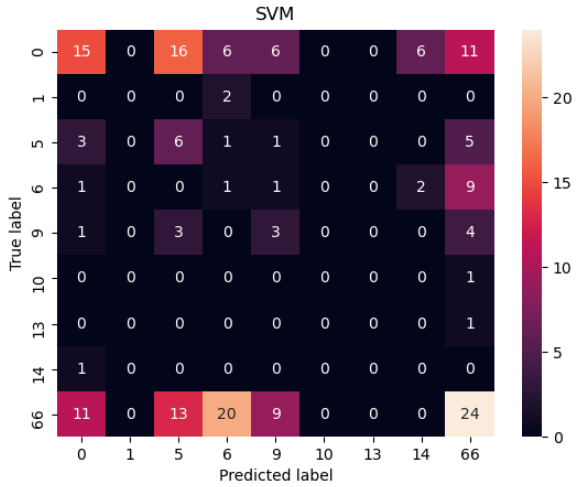
C.4 Including the data of the inter-annotation study



(a) Binary



(b) Combined



(c) All

Figure C.1: The confusion matrices of the best performing models for the different labelsets trained and evaluated on the data of the inter-annotation study.

Appendix D

Relevant documents

D.1 Spirometry attempts assessing form

This form is used by the professionals to file the errors found during the spirometry attempts of the tests performed during the short-term study, described in section 6.1.1, and the inter-annotation study, explained in section 6.2.1.

	Manoeuvre 1	Manoeuvre 2	Manoeuvre 3	Manoeuvre 4	Manoeuvre 5	Manoeuvre 6	Manoeuvre 7	Manoeuvre 8
Wrong posture								
Unsatisfactory start								
Flow leak								
Obstructive mouthpiece								
No maximal effort								
Cough								
Unsatisfactory end								
Extra breath								
Other, namely:								

General remarks:

D.2 Letter for the participants of the inter-annotation study

This letter was read by the participants of the inter-annotation study before the spirometry tests were executed. The email addresses and phone numbers are covered for privacy reasons.

Beste,

Mijn naam is Iris Heerlien en voor de Universiteit Twente ben ik bezig met een afstudeeronderzoek naar het thuis monitoren van kinderen met astma. Dit onderzoek is onderdeel van het project SPIROmetry-based PLAYful Asthma (SpiroPlay). Het thuismonitoren van astma houdt in dat de kinderen thuis een spirometrie test kunnen doen, zonder dat hier een arts of vakkundige bij nodig is. Een spirometrie test is een blaastest waarbij zo hard mogelijk, en zo lang mogelijk uit geblazen moet worden, zodat hun long capaciteit gemeten kan worden. Als de kinderen thuis gemonitord kunnen worden, in plaats van vaak naar het ziekenhuis te moeten, geeft dit de kinderen een enorme vrijheid!

Een onderdeel van dit onderzoek is zorgen dat de fouten, die gemaakt kunnen worden tijdens een blaastest, automatisch worden gedetecteerd. Voorbeelden van fouten zijn: hoesten tijdens de meting, niet hard genoeg uitademen, of niet lang genoeg uitademen. Om deze fouten te detecteren, maak ik een computer programma. Om te weten hoe goed mijn computer programma deze fouten kan herkennen, moet ik een aantal blaastesten afnemen om deze door vakkundigen te laten beoordelen. Dit moet ik doen om te kijken hoe goed de vakkundigen de fouten erin kunnen herkennen. Als ik dit weet, kan ik beoordelen hoe goed mijn computer programma dit kan in vergelijking met de vakkundigen.

De vakkundigen bepalen welke fouten er in een blaastest zit aan de hand van de grafiek die gegenereerd wordt uit de data van de blaastest, en door te kijken naar de persoon die de blaastest doet.

Deze blaastest zou ik graag bij u willen afnemen in een openbare ruimte. Dit houdt in dat, als u akkoord gaan met de deelname aan dit onderzoek, u wordt gevraagd om de blaastest 2 keer uit te voeren, één keer met een app met metaforen die u sturing geeft tijdens uw test, en één keer met sturing gegeven door de onderzoeker. Elke test is 3 tot 8 metingen, afhankelijk van of de metingen wel of geen fouten bevatten. Het apparaatje wat hiervoor gebruikt wordt ziet er zo uit:

Een voorbeeld van een metafoor in de app is dat u tijdens de meting een auto moet laten rijden. Dit ziet er zo uit:

Als u inademt, dan gaat de toerenteller oplopen. Als u uitademt, gaat de auto rijden en veranderd in een sportauto.

Deze blaastesten duren in totaal, met uitleg, 10 tot 15 minuten. We voorzien hier geen risico's bij. Er is geen financiële compensatie, maar we verwachten wel dat het leuk is om deel te nemen. Van te voren wordt u verteld hoe u moet blazen. Tijdens de blaastest zou ik u ook graag willen filmen. De vakkundige hoeft dan zelf niet aanwezig te zijn tijdens de meting , maar hij heeft deze informatie nodig om te kijken of, en welke, fout(en) er in de blaastest zit(ten). Volgens de standaard van onderzoek doen in Nederland (VSNU) worden deze video's 10 jaar bewaard. De video's blijven alleen toegankelijk voor de mensen die direct bij het onderzoek betrokken zijn. Ook zullen de video's nergens gepubliceerd worden

Tijdens het onderzoek wordt u op geen enkele manier gediagnostiseerd met het wel of niet hebben van astma. Hiervoor zijn de opzet en resultaten niet geschikt.

Via deze brief vraag ik uw toestemming om deel te nemen aan dit onderzoek. Het meedoen is geheel vrijwillig en niet meedoen zal geen verdere gevolgen hebben. U mag op ieder moment tijdens het experiment aangegeven dat u wilt stoppen, zonder een reden te hoeven geven. Aan meedoen zijn geen kosten verbonden.

Als u ermee akkoord gaat dat u meedoet aan deze blaastesten, dan wil ik u vragen om het bijgevoegde toestemmingsformulier positief in te vullen. Voor meer informatie kunt u contact opnemen met de betrokken onderzoekers, Iris Heerlien en Robby van Delden, via de contact informatie onder aan deze brief. Mocht u onafhankelijk advies willen, dan kunt u contact opnemen met de onafhankelijke ethische commissie van de universiteit Twente (ethics-comm-ewi@utwente.nl).

Bij voorbaat dank,

Met vriendelijke groet,

Iris Heerlien

Afstudeer student masters 'Data Science' en 'Human Media Interaction' aan de Universiteit Twente

Contact informatie:

E-mail:

Telefoon nummer:

Robby van Delden

Universitair Docent Human Media Interaction, aan de Universiteit Twente

Contact informatie:

E-mail:

Telefoon nummer:

D.3 Consent form for the participants of the inter-annotation study

This consent form was signed by the participants of the inter-annotation study. The email addresses and phone numbers are covered for privacy reasons.

Toestemmingsformulier Inter-annotatie studie SpiroPlay
Betreft: Toestemming voor deelname aan blaastesten voor onderzoek naar thuis monitoren van astma van de Universiteit Twente.

Als u akkoord gaat dat u meedoet kunt u hieronder aankruisen dat u toestemming geeft, de verdere gegevens invullen en het formulier ondertekenen.

☐ **Ik ben over dit onderzoek volledig geïnformeerd en geef toestemming dat ik hieraan deelneem. Ik ben me ervan bewust dat deelname geheel vrijwillig is. Ik heb de mogelijkheid gehad om vragen te stellen aan de betrokken onderzoekers ( ) of aan de onafhankelijk ethische commissie (ethics-comm-ewi@utwente.nl) en eventuele vragen zijn beantwoord. Ik geef toestemming voor het verzamelen van data en onderzoeksmaterialen zoals beschreven in de bijbehorende brief. Ik geef ook toestemming voor het maken van video-opnames. De video's worden enkel door betrokken onderzoekers bekeken en zullen nooit publiek worden gemaakt of vertoond aan derden voor demonstratie of rapportage.**

☐ **Ik ben mij ervan bewust dat deze activiteit niet gebruikt wordt voor diagnostische doeleinden, aangezien de huidige opzet en resultaten hier niet geschikt voor zijn. Ik ben mij ervan bewust dat er door de aard van de studie geen uitspraak kan worden gedaan over het al dan niet hebben van astma.**

Naam……

Datum…….

Handtekening participant: …….

Als onderdeel van de Universiteit Twente zijn we verplicht de Verordering Algemene Gegevensbescherming (AVG) en Uitvoeringswet na te leven. We hanteren hiervoor maatregelen met betrekking tot verwerking en inzage van persoonlijk identificeerbare data, zoals namen, video, foto's, en geluidsopnames**.**

Contact informatie Mocht u vragen hebben over dit onderzoek dan kunt u contact opnemen met Iris Heerlien ( ), Robby van Delden ( ), en voor een onafhankelijk advies (ethics-comm-ewi@utwente.nl).

D.4 Processing form for the raters of the inter-annotation study

This form was signed by the raters of the inter-annotation study. The email addresses and phone numbers are covered for privacy reasons.

Verwerkingsformulier Inter-annotatie studie SpiroPlay
Betreft: Het akkoord verwerken van gegevens verworven tijdens het experiment voor de inter-annotatiestudie voor het project SpiroPlay.

Als u akkoord gaat met de verwerkingseisen kunt u hieronder aankruisen dat u zich eraan zult houden, de verdere gegevens invullen en het formulier ondertekenen.

☐ **Ik begrijp dat de data die ik ontvang persoonlijke data is en zal het niet verstrekken aan andere partijen.**

☐ **Ik zal de data verwijderen zodra ik klaar ben met de verwerking ervan.**

☐ **Ik zal de data niet gebruiken voor persoonlijke doeleinden**

Naam……

Datum……..

Handtekening : ……..

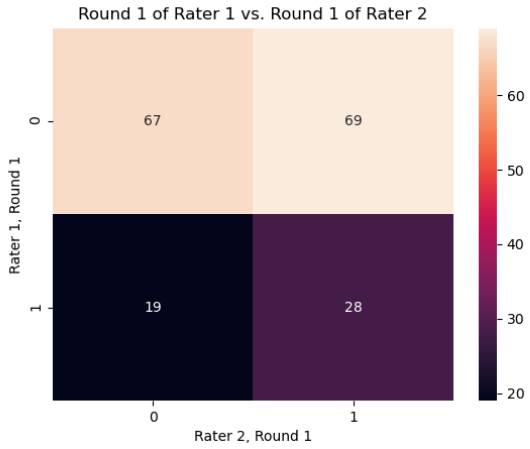
Contact informatie Mocht u vragen hebben over dit onderzoek en/of de verwerkingseisen dan kunt u contact opnemen met Iris Heerlien ( ), Robby van Delden ( ), en voor een onafhankelijk advies (ethics-comm-ewi@utwente.nl).

Appendix E

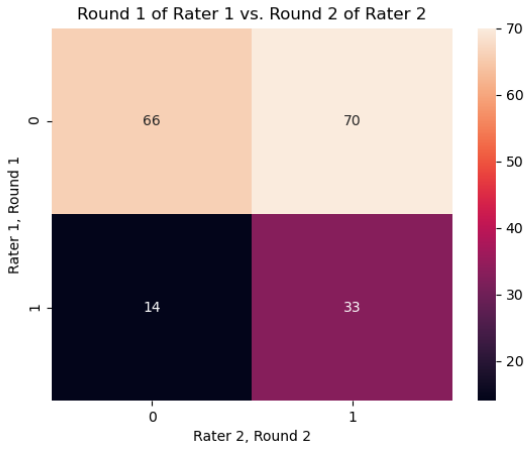
Results inter-annotation study

E.1 Labelset: Binary

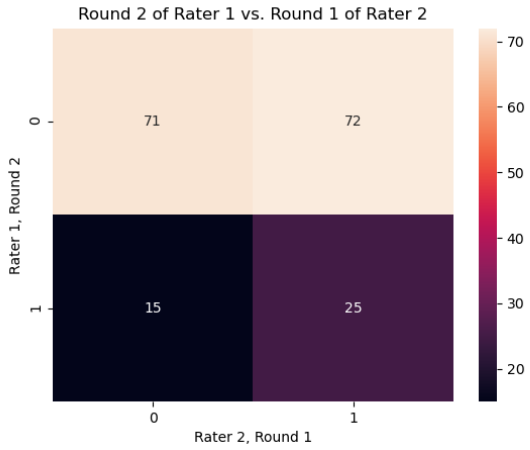
E.1.1 Inter-rater agreement



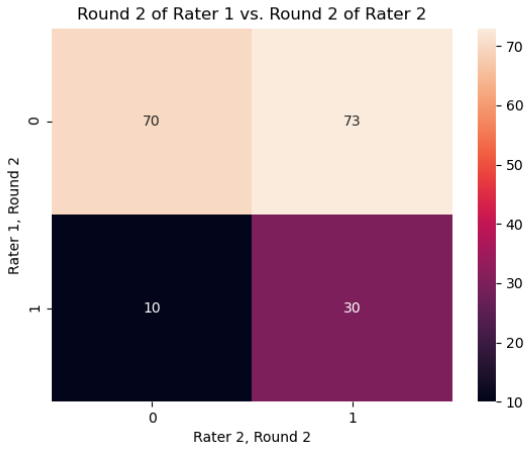
(a) Round 1 of Rater 1 vs. Round 1 of Rater 2



(b) Round 1 of Rater 1 vs. Round 2 of Rater 2

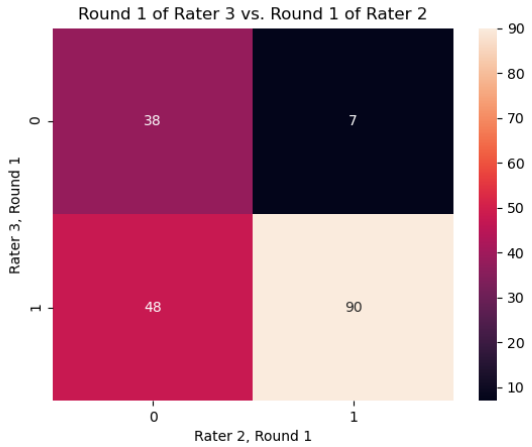


(c) Round 2 of Rater 1 vs. Round 1 of Rater 2

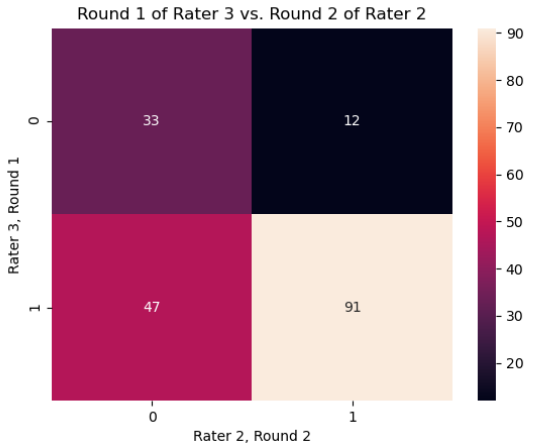


(d) Round 2 of Rater 1 vs Round 2 of Rater 2

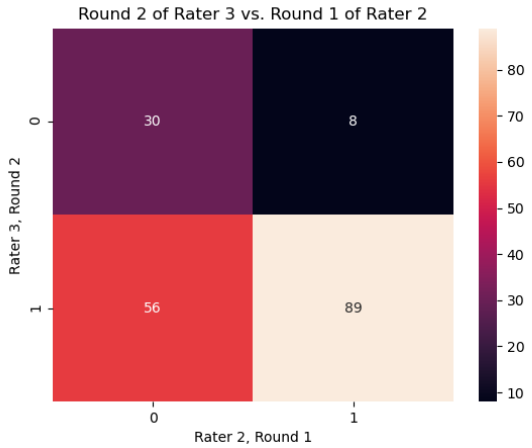
Figure E.1: The confusion matrices comparing the labels given by the raters 1 and 2 during the two rounds, using labelset 'binary'.



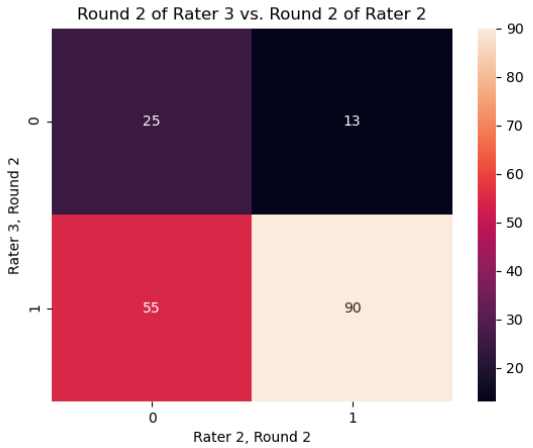
(a) Round 1 of Rater 3 vs. Round 1 of Rater 2



(b) Round 1 of Rater 3 vs. Round 2 of Rater 2

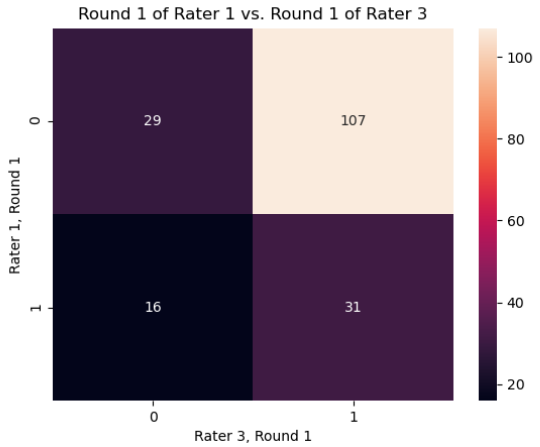


(c) Round 2 of Rater 3 vs Round 1 of Rater 2

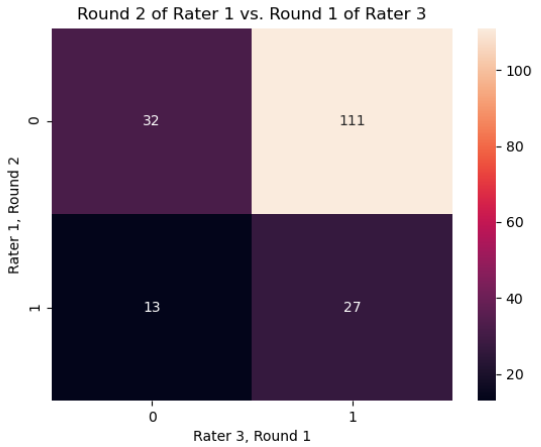


(d) Round 2 of Rater 3 vs Round 2 of Rater 2

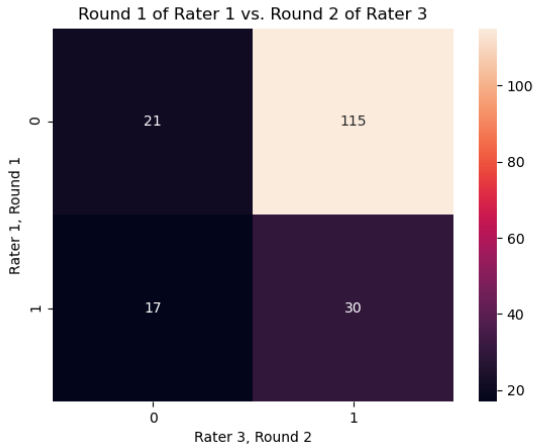
Figure E.2: The confusion matrices comparing the labels given by the raters 2 and 3 during the two rounds, using labelset 'binary'.



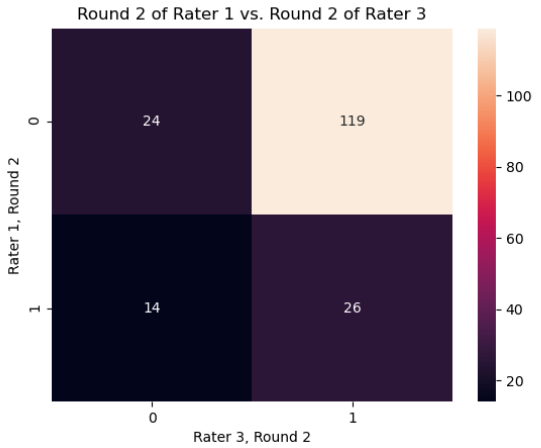
(a) Round 1 of Rater 1 vs. Round 1 of Rater 3



(b) Round 2 of Rater 1 vs. Round 1 of Rater 3



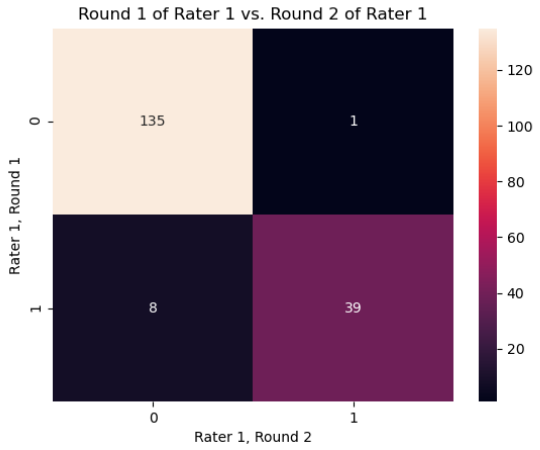
(c) Round 1 of Rater 1 vs Round 2 of Rater 3



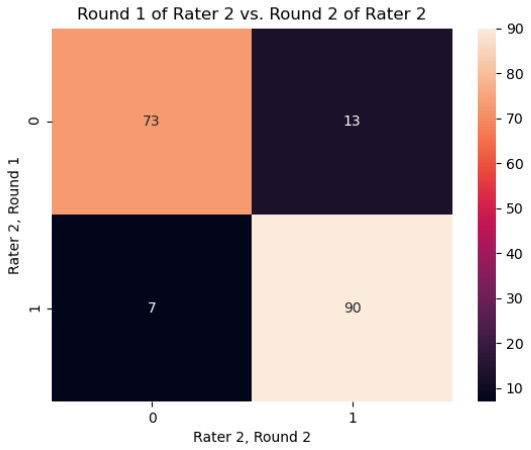
(d) Round 2 of Rater 1 vs Round 2 of Rater 3

Figure E.3: The confusion matrices comparing the labels given by the raters 1 and 3 during the two rounds, using labelset 'binary'.

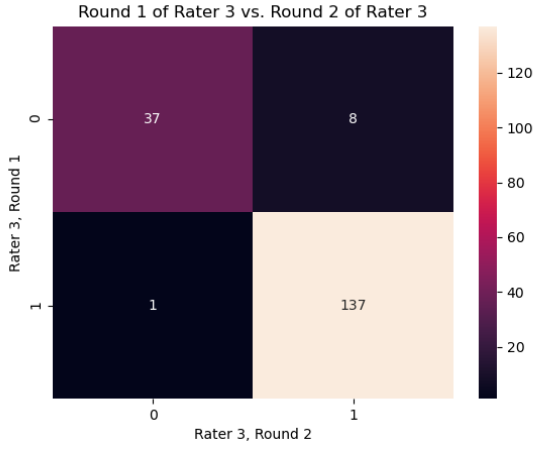
E.1.2 Intra-rater agreement



(a) Round 1 vs. Round 2 of Rater 1



(b) Round 1 vs. Round 2 of Rater 2

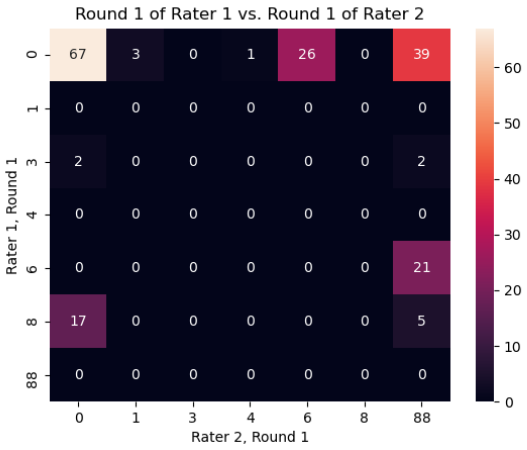


(c) Round 1 vs. Round 2 of Rater 3

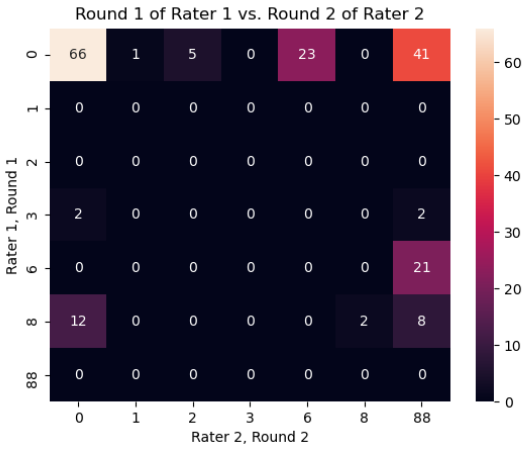
Figure E.4: The confusion matrices comparing the labels given by the two raters during the first and second round, using labelset 'binary'.

E.2 Labelset: Combined

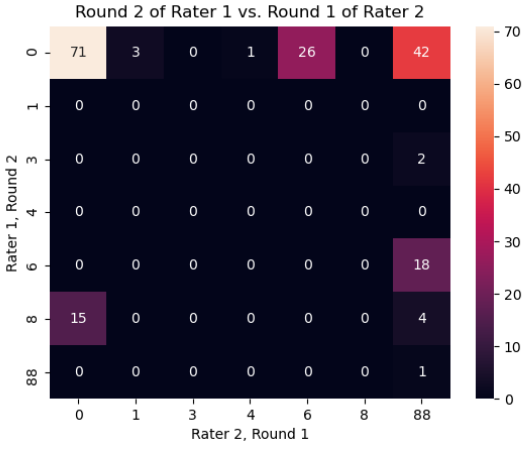
E.2.1 Inter-rater agreement



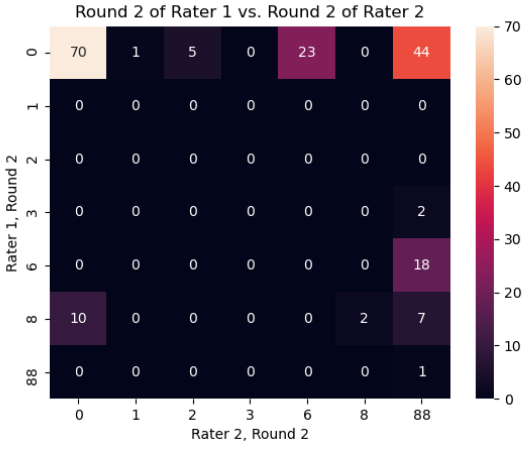
(a) Round 1 of Rater 1 vs. Round 1 of Rater 2



(b) Round 1 of Rater 1 vs. Round 2 of Rater 2

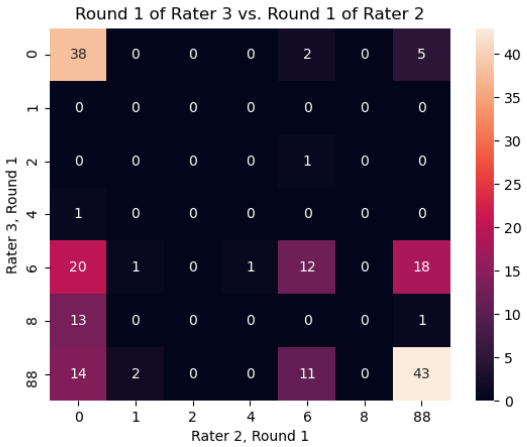


(c) Round 2 of Rater 1 vs. Round 1 of Rater 2

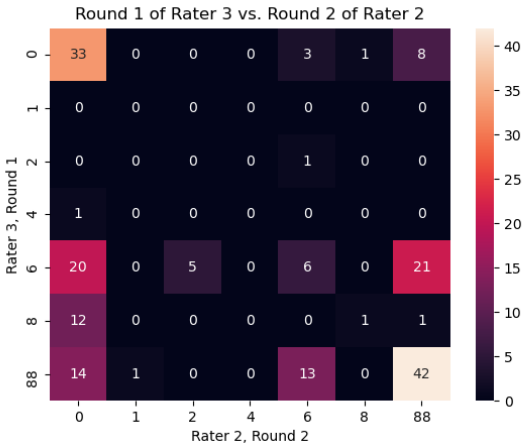


(d) Round 2 of Rater 1 vs Round 2 of Rater 2

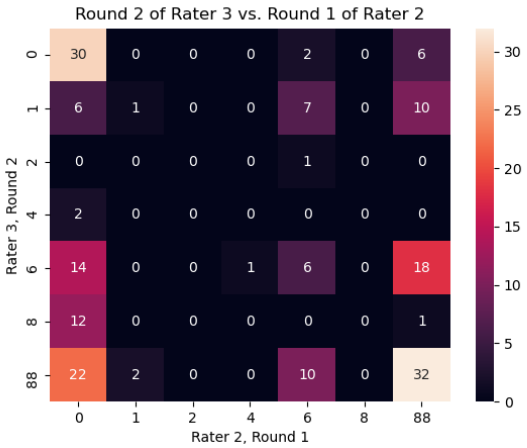
Figure E.5: The confusion matrices comparing the labels given by the raters 1 and 2 during the two rounds, using labelset 'combined'.



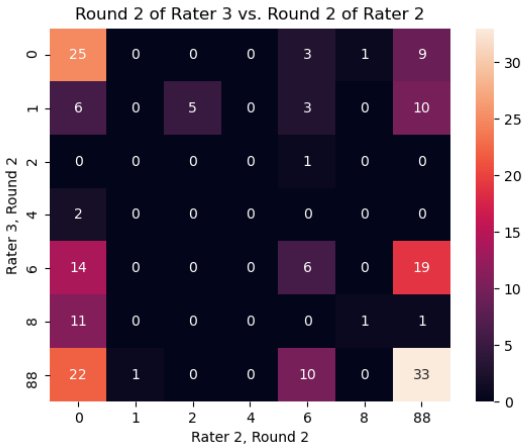
(a) Round 1 of Rater 3 vs. Round 1 of Rater 2



(b) Round 1 of Rater 3 vs. Round 2 of Rater 2

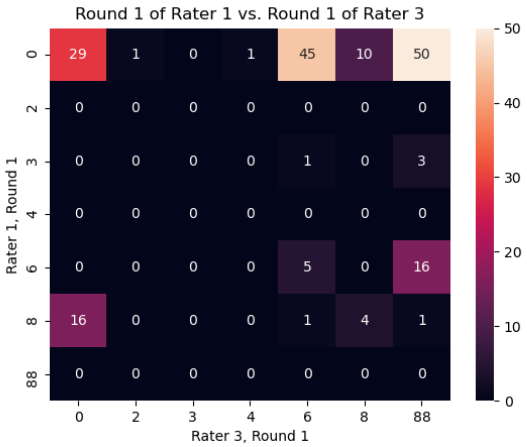


(c) Round 2 of Rater 3 vs. Round 1 of Rater 2

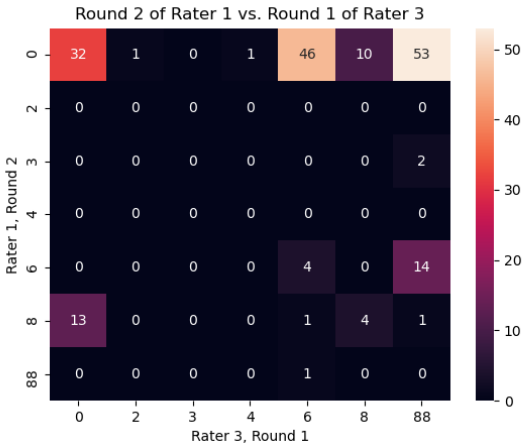


(d) Round 2 of Rater 3 vs. Round 2 of Rater 2

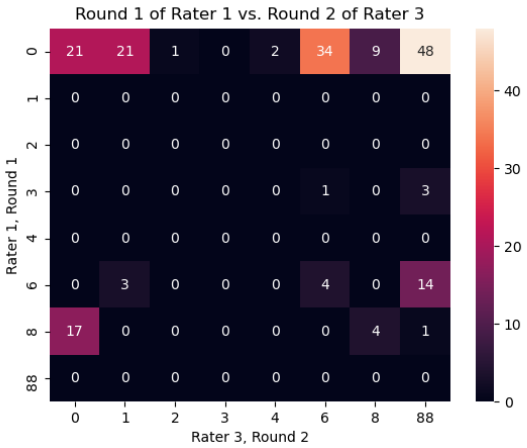
Figure E.6: The confusion matrices comparing the labels given by the raters 2 and 3 during the two rounds, using labelset 'combined'.



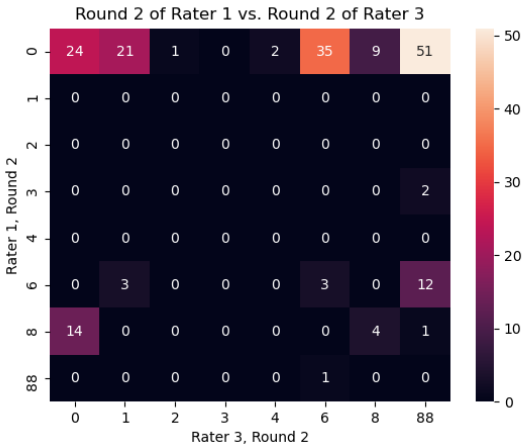
(a) Round 1 of Rater 1 vs. Round 1 of Rater 3



(b) Round 2 of Rater 1 vs. Round 1 of Rater 3



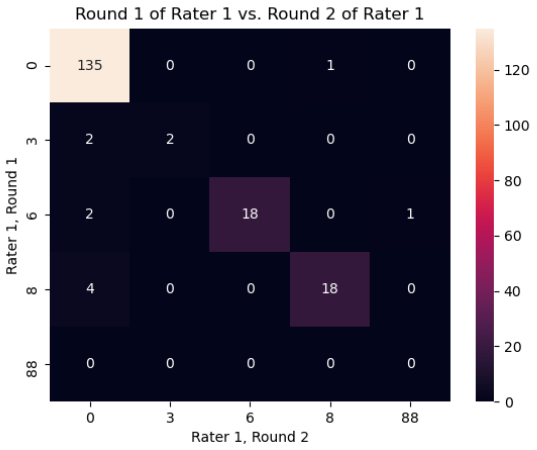
(c) Round 1 of Rater 1 vs. Round 2 of Rater 3



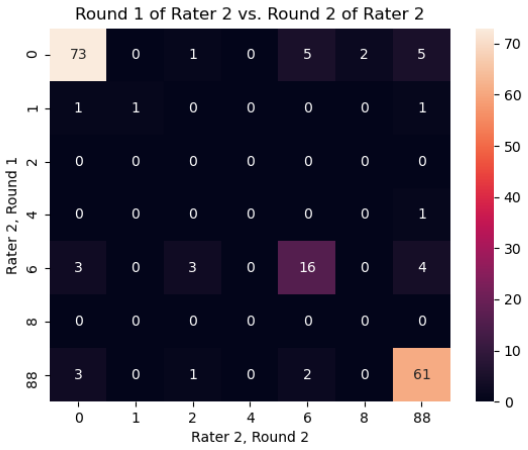
(d) Round 2 of Rater 1 vs. Round 2 of Rater 3

Figure E.7: The confusion matrices comparing the labels given by the raters 1 and 3 during the two rounds, using labelset 'combined'.

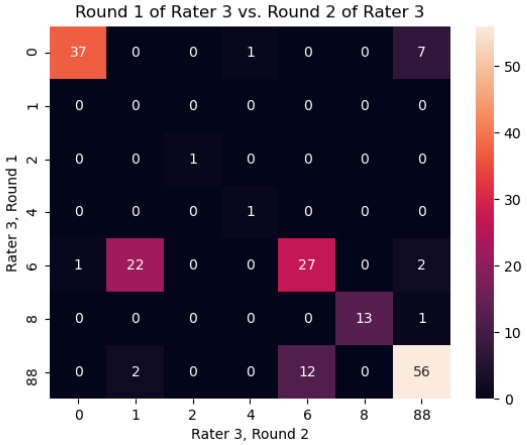
E.2.2 Intra-rater agreement



(a) Round 1 vs. Round 2 of Rater 1



(b) Round 1 vs. Round 2 of Rater 2

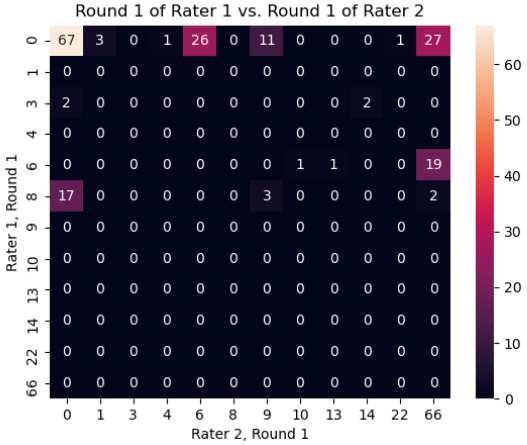


(c) Round 1 vs. Round 2 of Rater 3

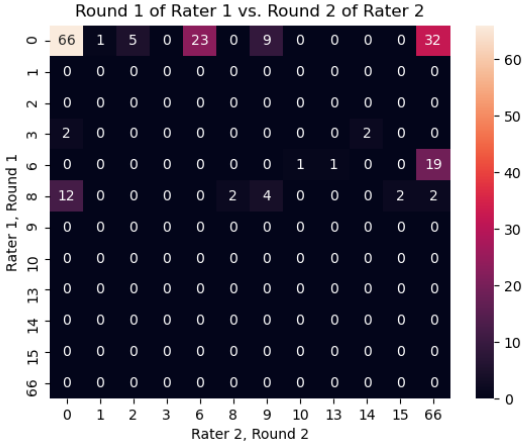
Figure E.8: The confusion matrices comparing the labels given by the two raters during the first and second round, using labelset 'combined'.

E.3 Labelset: All

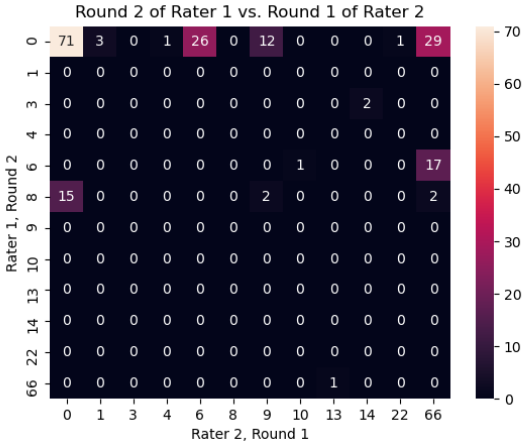
E.3.1 Inter-rater agreement



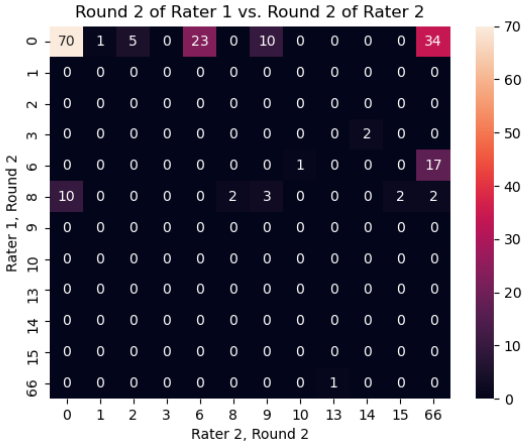
(a) Round 1 of Rater 1 vs. Round 1 of Rater 2



(b) Round 1 of Rater 1 vs. Round 2 of Rater 2

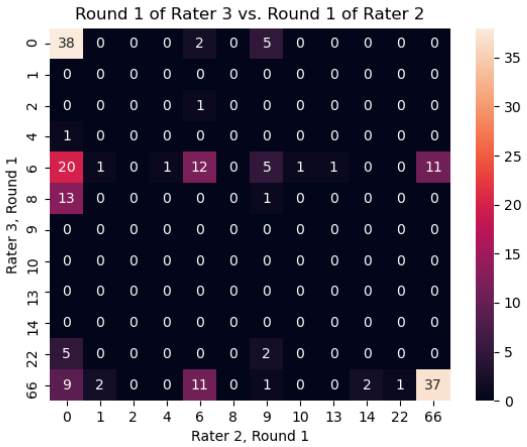


(c) Round 2 of Rater 1 vs. Round 1 of Rater 2

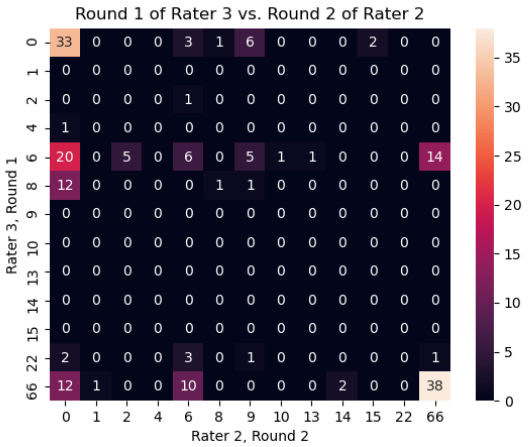


(d) Round 2 of Rater 1 vs Round 2 of Rater 2

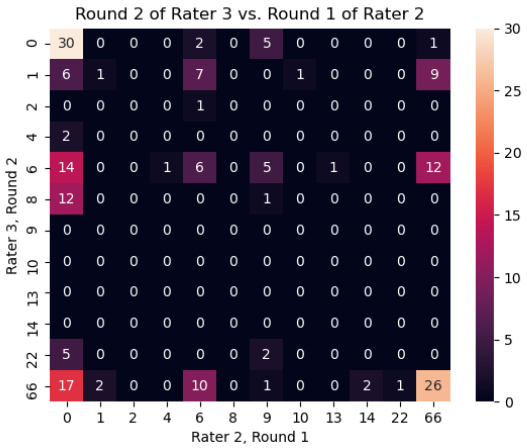
Figure E.9: The confusion matrices comparing the labels given by the raters 1 and 2 during the two rounds, using labelset 'all'.



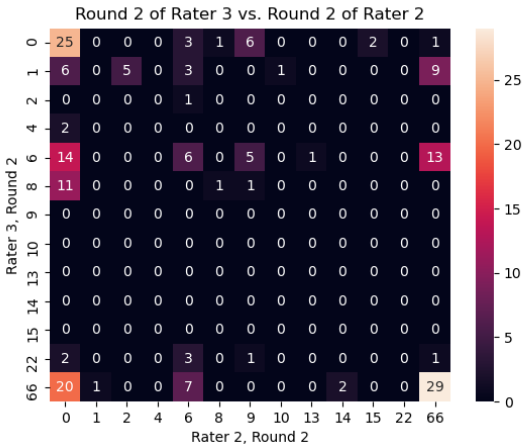
(a) Round 1 of Rater 3 vs. Round 1 of Rater 2



(b) Round 1 of Rater 3 vs. Round 2 of Rater 2

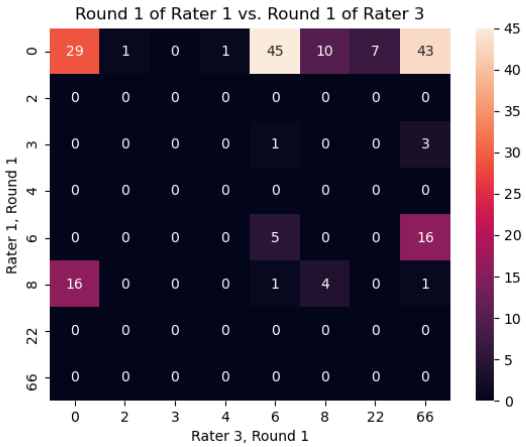


(c) Round 2 of Rater 3 vs. Round 1 of Rater 2

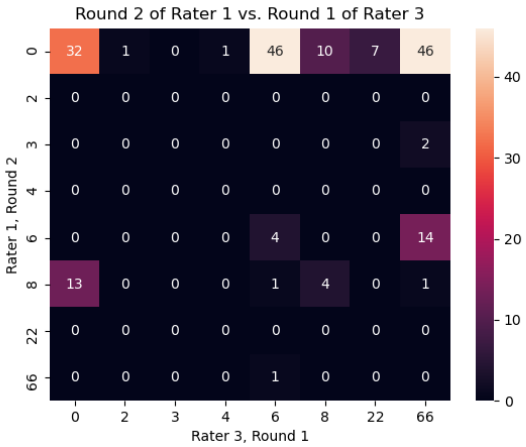


(d) Round 2 of Rater 3 vs. Round 2 of Rater 2

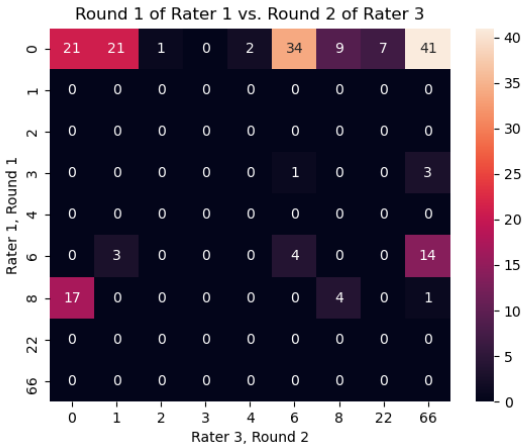
Figure E.10: The confusion matrices comparing the labels given by the raters 2 and 3 during the two rounds, using labelset 'all'.



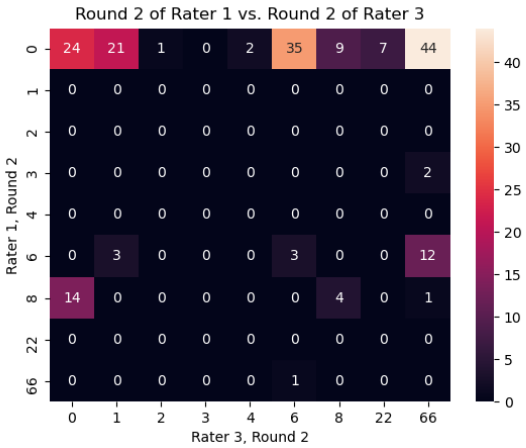
(a) Round 1 of Rater 1 vs. Round 1 of Rater 3



(b) Round 2 of Rater 1 vs. Round 1 of Rater 3



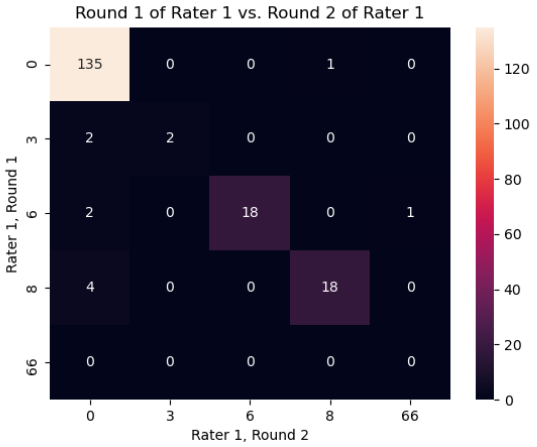
(c) Round 1 of Rater 1 vs. Round 2 of Rater 3



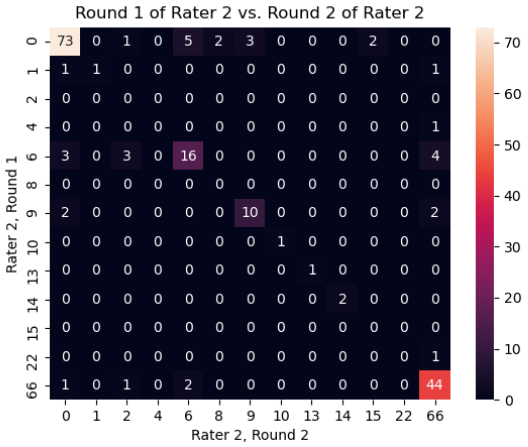
(d) Round 2 of Rater 1 vs. Round 2 of Rater 3

Figure E.11: The confusion matrices comparing the labels given by the raters 1 and 3 during the two rounds, using labelset 'all'.

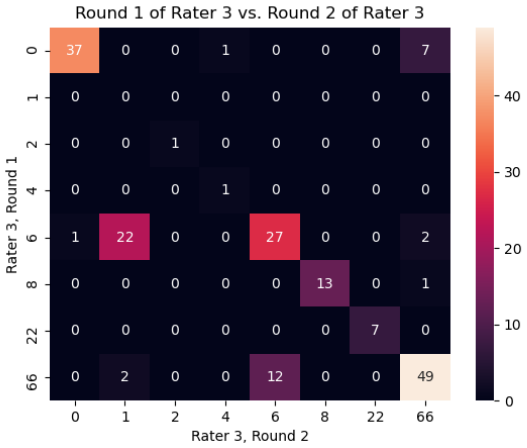
E.3.2 Intra-rater agreement



(a) Round 1 vs. Round 2 of Rater 1



(b) Round 1 vs. Round 2 of Rater 2



(c) Round 1 vs. Round 2 of Rater 3

Figure E.12: The confusion matrices comparing the labels given by the two raters during the first and second round, using labelset 'all'.