UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Richard Annilo

Predicting Respiratory Diseases from Lung Sounds Using Machine Learning

Bachelor's thesis (9 ECTS)

Supervisor: MSc. Dmytro Fishman

Tartu 2021

Predicting Respiratory Diseases from Lung Sounds Using Machine Learning

Abstract:

Respiratory diseases are a leading cause of death worldwide. Using machine learning for diagnosis could significantly reduce costs and radiation exposure due to X-ray and CT scans, and improve accessibility to places with limited technology or less-experienced staff. While similar technologies have been successfully applied in the medical field before, sound signal analysis is still in its early stages with significant potential.

This thesis's goal was to create a codebase to help researchers enter and advance the field of respiratory sound analysis. In total, six experiments were conducted with four classical machine learning and one deep learning algorithm. The aim was to classify six classes (five respiratory diseases and one class for healthy patients) using a database of respiratory sounds and patient data. Test results, which used macro-averaged F1-scores as the primary evaluation metric, showed that SVM and decision tree models worked best (scores 0.62 and 0.54), while the convolutional neural network models performed worst (best score 0.3). The differences in the models' performances were most likely affected by the dataset's noisiness and umbalancedness. Further research and better data would be required for any conclusive results.

The source code for this thesis is publicly available in a Github repository [1].

Keywords: Machine learning, deep learning, audio signal analysis, respiratory diseases

CERCS: P170 Computer science, numerical analysis, systems, control

Kopsuhelide kasutamine hingamisteede haiguste ennustamiseks masinõppe abil

Lühikokkuvõte:

Hingamisteede haigused on kogu maailmas üks peamiseid surma põhjustajaid. Masinõppe kasutamine diagnoosimiseks võib oluliselt vähendada röntgen- ja kompuutertomograafia tõttu tekkivaid kulusid ja kiiritust, samuti parandada ligipääsu piiratud tehnoloogiaga või vähem kogenud personaliga kohtadele. Kuigi sarnaseid tehnoloogiaid on meditsiinivaldkonnas varemgi edukalt rakendatud, on helisignaalide analüüs endiselt varases staadiumis ning märkimisväärse potentsiaaliga.

Selle lõputöö eesmärk oli luua koodibaas, mis aitaks teadlastel siseneda ja edendada hingamisteede heli analüüsi valdkonda. Kokku viidi läbi kuus katset nelja klassikalise masinõppe ja ühe süvaõppe algoritmiga. Eesmärk oli klassifitseerida kuus klassi (viis hingamisteede haigust ning üks tervete patsientide klass), kasutades andmebaasi hingamisteede helidest ja patsientide andmetest. Testitulemused, milles põhilise hindamismõõdikuna kasutati makro-keskmistatud F1-skoori, näitasid, et kõige paremini töötasid SVM ja otsustuspuu mudelid (hinded 0,62 ja 0,54), halvemini konvolutsioonilise närvivõrgu (CNN) mudelid (parim tulemus 0,3). Mudelite jõudluse erinevusi mõjutas tõenäoliselt mürased andmete ja klasside erinevad andmemahud. Lõplike tulemuste saamiseks oleks vaja täiendavaid uuringuid ja paremaid andmeid.

Lõputöö lähtekood on avalikult kättesaadav Github'i repositooriumis [1].

Võtmesõnad: Masinõpe, süvaõpe, helitöötlus, kopsuhaigused

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Table of Contents

1 Introduction

Respiratory diseases account for more than 4 million premature deaths each year [2]. After cardiovascular diseases, they are the largest contributors to the global disease burden [2]. The recent coronavirus pandemic has further brought into public discussion the importance of accessible and accurate respiratory disease diagnosis.

Currently, chest X-ray and computed tomography (CT) scans are often used to diagnose patients. These scans, however, are costly and increase the risk of cancer due to radiation [3]. Stethoscopes are also used to listen to the patient's lungs, but it is not enough to make a reliable diagnosis [4]. Using machine learning (ML) to diagnose patients would mitigate radiation risks and decrease costs. It would also make accurate predictions more accessible to developing countries or remote geographical locations—places that do not have experienced staff or necessary funds for expensive equipment.

There is reason to believe this could be possible. Machine learning has already been successfully applied in various medical contexts, such as detecting heart disease [5] and using medical imagery to find signs of skin cancer [6] or tuberculosis [7]. While these have often resulted from advanced computer vision technologies, lung sound analysis is still in its early stages, and existing small-scale studies show great potential [8]. Lung disease classification algorithms still require much further research and attention to reach wide use in the clinical setting.

The purpose of the thesis is to boost the field of lung sound analysis by creating and sharing a codebase that includes experiments with various machine learning models on a database of lung sounds. While similar research has mostly focused on binary classification, this thesis will classify six different classes (five respiratory diseases and one class for "healthy"). Six experiments will be conducted: 3 of which with four machine learning models and three on a convolutional neural network (CNN).

The "Background" chapter describes the relevant medical and sound signal background. The machine learning models used in this thesis, information about the dataset, evaluation methods, frameworks, and computational resources are described in "Methods". Specifics about preprocessing and training for the machine learning models are described in "Experiments". In "Results", the findings of the previous chapter are described and analysed. The "Conclusions" chapter summarises the essential findings and lists ideas for future work.

2 Background

This chapter describes the relevant background information to this thesis in two parts. First, the medical background focuses on the human respiratory system under normal conditions and during infection. The sound signal background chapter describes how sound works and how sound features can be extracted for machine learning purposes.

2.1 Medical background

Breathing is common to all humans. However, most of it happens without our conscious awareness or control. To understand lung sounds and lung diseases, it is essential to know how the human respiratory system works under normal conditions. This chapter will explain that, followed by a description of how the system can malfunction and cause abnormal sounds. Finally, all the respiratory diseases covered in this thesis are introduced.

2.1.1 The human respiratory system

The human body's respiratory system is responsible for getting oxygen into the body and getting carbon dioxide out. This process is known as breathing. When a person breathes in, air particles move from their nose to their lungs. Inside the lungs, there is a network of branching airways (Figure 1).

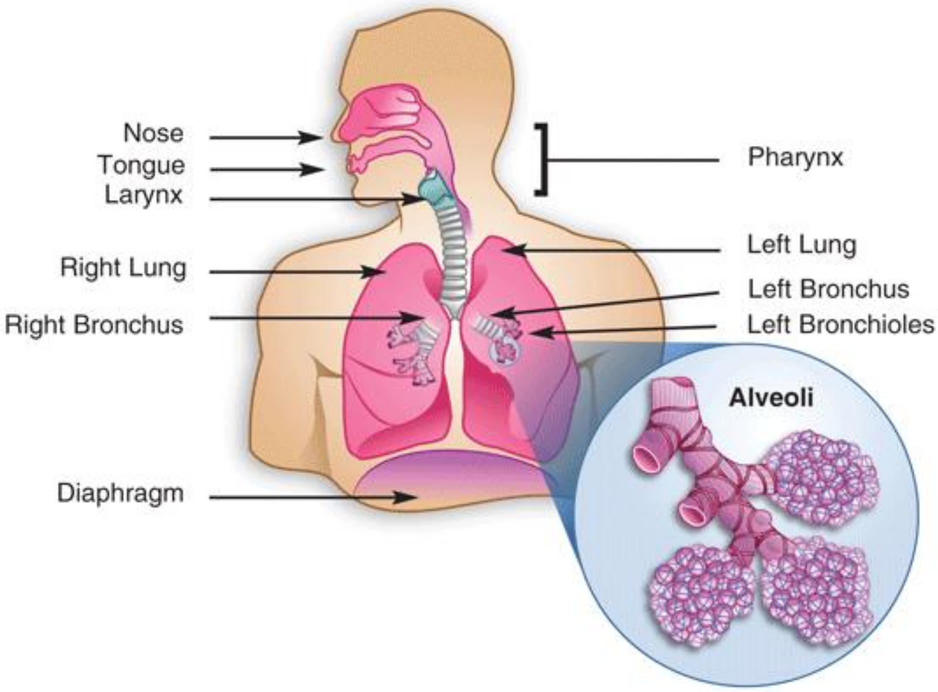


Figure 1. Simplified diagram of the human respiratory system [9].

This network begins with the trachea and branches into smaller and smaller segments (bronchi and bronchioli), ending with the alveoli. These small and hollow alveoli are responsible for exchanging oxygen and carbon dioxide, the respiratory system's primary function.

2.1.2 Crackles and wheezes

Under normal conditions, air movement in the lungs produces no noticeable sounds. However, when foreign particles (tobacco smoke) or pathogens (fungi, bacteria, viruses, parasites) infect or cause damage to a part of the respiratory system, the immune system sends out a response with the intention of neutralising the threat. This response results in noticeable symptoms in the affected area, such as a congested nose, sore throat, or abnormal lung sounds.

There are many types of abnormal (adventitious) lung sounds, two of which are crackles and wheezes. Crackles can be heard when air movements open blocked airways or make liquids in the alveoli bubble [10]. They can be described as discontinuous crackling or rattling sounds (Figure 2). On the other hand, wheezes are longer in duration and have a whistling quality [10] (Figure 3). They occur when airways in the lung are contracted or when air paths are obstructed [10]. These sounds are mostly not present in healthy people. Thus they are abnormal.

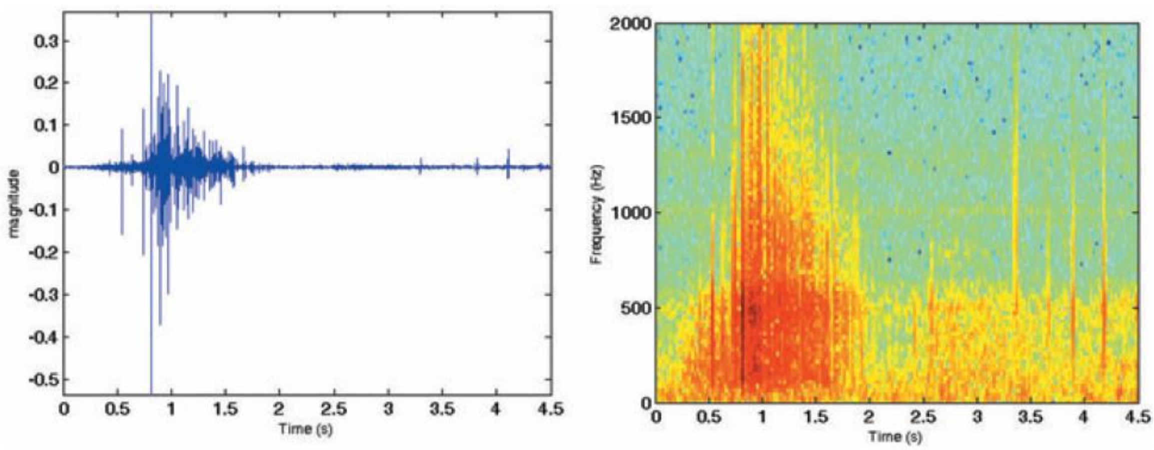


Figure 2. Crackles in time and time-frequency domains [10].

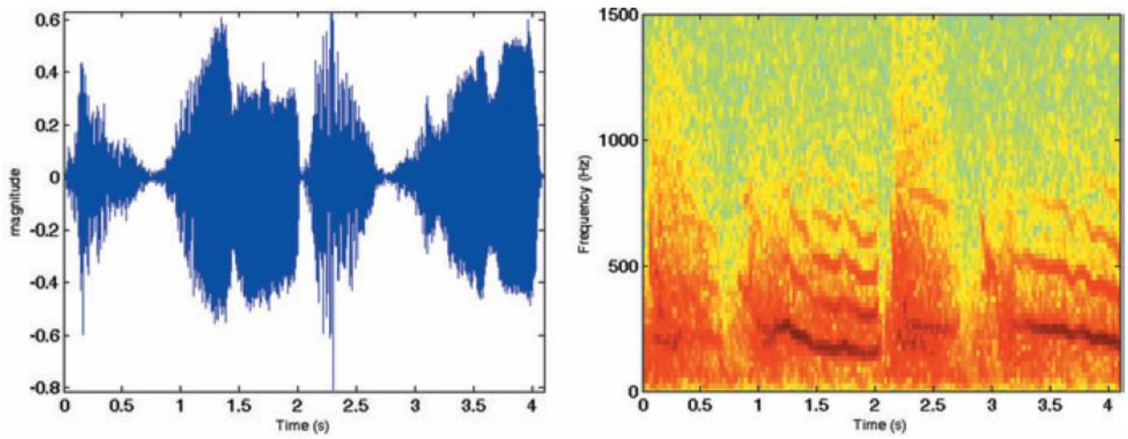


Figure 3. Wheezes in time and time-frequency domains [10].

Both wheezes and crackles are easily detected by listening to the patient's lungs with a stethoscope [10]. However, knowing that a patient has these sounds is not enough to know what treatment is necessary. Minute differences in the coarseness of crackles or wheezes' intensity can influence the disease's severity and, thus, the correct approach to curing it [10]. That is why diagnosing patients by their lung sounds is often complicated.

2.1.3 Respiratory diseases

The aforementioned abnormal sounds can be caused by various respiratory diseases [11], but only five are relevant to this thesis. These five are listed below.

1. If the upper part of the respiratory system (from the nose to the larynx) is infected, it causes an illness called the **upper respiratory tract infection (URTI)** [11]. This is often referred to as the common cold. URTI mostly causes mild symptoms, like a sore throat or a blocked nose [11]. It does not cause wheezing or crackles—these sounds are related to the lower part of the respiratory system.
2. When the lung's large airways (major bronchi) are infected, the resulting condition is called **bronchiectasis** [11]. Due to the immune system's attempt at fighting the infection, the bronchi are inflamed with mucus and have smaller air paths than usual, causing wheezing [10]. A high-resolution computer tomography (HRCT) of the patient's chest is required to diagnose the patient correctly [4]. This requires exposing the patient to X-radiation.
3. **Bronchiolitis** is an infection of the lungs' small airways (bronchioles) [11]. It happens mostly to young children under the age of 2, and, similarly to bronchiectasis, it can cause wheezing because the airways in the lungs are inflamed [11].

4. In the case of **pneumonia**, the infection has gotten to the very end of the air paths: bronchioles and alveoli [11]. This causes the immune system to fight the disease by causing inflammation and partially filling the alveoli with secretion, thus obstructing normal functioning and causing crackling sounds when breathing [11]. Wheezing can also occur, though it is less common [12].
5. **Chronic obstructive pulmonary disease (COPD)** is different from the others because it is caused by tobacco smoke, air pollution, toxic fumes, or other foreign particles rather than pathogens [11]. It is a common disease and a general term used to characterise many different conditions. Because it usually causes inflammation in some parts of the lung, wheezing can be heard [13].

Next, an overview of sound and useful sound features is provided.

2.2 Sound signal background

To predict respiratory diseases, sounds can be used to identify and categorise them. This chapter describes how sound works, lists some common methods for visualising sound, and introduces the sound features used in this thesis.

2.2.1 Time-amplitude domain

When a person claps, the air pressure around their hands changes and ripples outward as waves. When these waves reach their ears, it is perceived as sound. If these waves reach a microphone, they are converted into changes in voltage [14]. This produces a digital sound file and can be represented as a time-amplitude graph called a waveform.

A waveform sound file can be thought of as a list of numbers, each number describing the amplitude of the soundwave at a certain point in time. The more numbers (more commonly "samples") there are describing these waves, the more information there is about the sound.

2.2.2 Time-frequency domain

The waveform of a sound only describes the amplitude over time. However, the Fourier transform (FT) applied to the waveform gives information about frequencies and their intensities (in dB, for instance) [14]. If the FT is used multiple times on each small segment of the original sound, the resulting matrix also gives information about how frequencies change over time. This matrix is called a spectrogram [14]. Because the result is a 2-dimensional array, much like an image, it can be used on deep learning models specialised for image recognition, such as a convolutional neural network.

However, the sound representation described above, in which frequencies range from very low to too high, may be redundant. For example, the human ear is more evolved to distinguish lower frequencies than higher ones better. For this reason, it often would not make sense to include many high frequencies in the spectrogram [15]. Therefore, the logarithm representation of a spectrogram (log-spectrogram) is often used.

Alternatively, the mel-frequency cepstral coefficients (MFCCs) can also be used. Just like spectrograms, they give information about frequencies and their intensities in time. However, MFCCs are distinguished by their feature ranges. Instead of spectrograms, MFCCs are designed to imitate the human ear's functioning even closer than log-spectrograms [15]. The lower frequency ranges increase linearly, and higher ranges increase logarithmically [15].

2.2.3 Additional sound features

Numerous features can be extracted from a sound. The following features were used in this thesis:

1. Spectral entropy indicates the uniformity or randomness of frequencies present in the sound [16].
2. Spectral rolloff describes the frequency below which some percentage of spectral energy (intensities of frequencies) is contained [17]. For instance, spectral rolloff 85 is the frequency below which 85% of the sound intensity comes from.
3. Root mean square is a method of calculating the average intensity of a sound [18].
4. Spectral centroid has been shown to differentiate between bright and sharp timbre in sounds [14].
5. Zero crossing rate is the rate at which the waveform changes sign (crosses the 0 line) [14]. If the value is high, it can indicate that the sound contains many high frequencies or is very noisy [14].
6. Spectral flatness is a feature that describes how much the sound resembles white noise [19].

In the next chapter, the relevant machine learning models and implementation details are presented.

3 Methods

"Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience [21:3]." In this chapter, the five such methods are introduced in two separate sections. The first focuses on classical machine learning models. These are easily interpretable and relatively simple models that are widely used. A newer and more complex algorithm XGBoost is also included in this section because it is based on other classical models, like the decision tree (DT). The second section focuses on a popular and powerful deep learning-based method: the convolutional neural network. Additionally, this chapter describes the dataset, evaluation methods, frameworks, and computational resources used in this thesis.

3.1 Classical machine learning methods

Here, four classical machine learning models used in this thesis are reviewed: decision tree, random forest (RF), XGBoost, and support vector machine (SVM).

Decision tree is a popular machine learning model used for classification and regression, dating back to the 1960s [22]. It is a tree-based model, consisting of a root node, decision nodes, and leaf nodes. Classifying a data sample involves starting from the root node, moving through the if-else conditions in the decision nodes, and arriving at a leaf node, which represents the predicted class (Figure 4).

As seen in Figure 4, all the decision tree details can be easily interpreted and understood. This is a major advantage of the model, as it makes it easier to gather insights into the prediction process and find possible mistakes in the model. Decision trees are also said to be robust to outliers, skewed data, and missing values [22], which makes them perform well even with noisy data. Its disadvantages include its relative simplicity, as many newer machine learning algorithms provide major improvements. Additionally, it can easily over- or underfit, especially if there are not enough data points [22].

Decision trees have been used in sound analysis and medical research before, for instance, diagnosing patients based on their symptoms [23], classifying marine animal sounds [24], and classifying common cardiovascular conditions problems using hearth sounds [25].

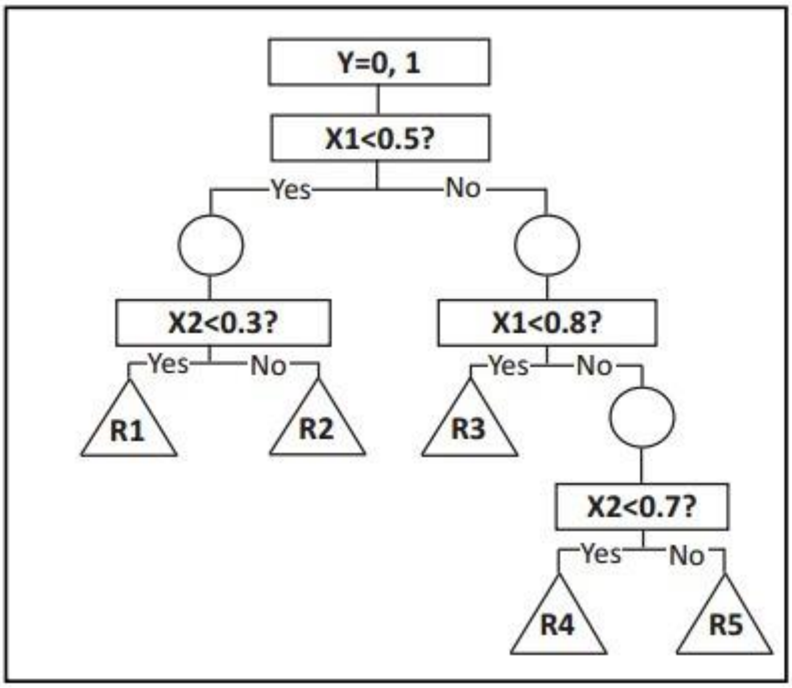


Figure 4. An example of a decision tree. The root node is “y=0,1”, and leaf nodes are named “R1” to “R5” [22].

Random forest is another popular machine learning method, first introduced in 1995 by Tim Kan Ho [26]. It works by creating multiple decision trees with a random subset of features and data rows. Outputs are the mean or the mode of those predictions. Compared to decision trees, a random forest model is less interpretable, as it may often consist of hundreds of individual trees. However, random forest models generally have higher performance and overfit less than individual decision trees [27].

XGBoost, initially released in 2014, is an efficient implementation of the gradient boosting method [28]. Gradient boosting is a machine learning technique which involves training multiple “weak” models (like decision trees) one by one sequentially. While training, the algorithm updates the weights of new models based on earlier models’ misclassified data points. It would be expected that XGBoost will outperform other tree-based models, such as decision trees and random forest because XGBoost has won many recent machine learning competitions [29] and its use of gradient boosting combines the prediction power of individual trees.

Support vector machine was chosen to contrast the other models, which are all based on decision trees. It can do binary classification by finding a hyperplane (a point for 2-D space, a line for 3-D space, etc.) at the maximal distance from both classes without being overly sensitive to outliers in the data. For data that is more difficult to classify (e.g., there are considerable overlaps between classes or the data cannot be linearly separated), SVMs can use an approach called the kernel trick, which maps the data into a higher-dimensional space without actually constructing it. The kernel trick makes it easier to separate even non-linearly separable data classes. A review article from 2018 notes that, together with neural networks, SVMs are one of the most commonly used methods to classify abnormal lung sounds [8].

Based on the complexity of the four models and performance reported in the previous studies, one might expect that the XGBoost or SVM will perform best in this thesis, followed by random forest, followed then by the decision tree.

3.2 Deep learning methods

Deep learning methods hold a lot of promise and have been shown to set new state-of-the-art results across various domains and problems, including sound classification. One of the most prominent deep learning methods used to date is the convolution neural network, which was also used in this thesis. It differs from other deep learning methods in its use of convolutional layers. The purpose of these layers is to find simple patterns of an input image (e.g., vertical lines) and then combine these patterns to recognise complex shapes such as the face of a dog or a person (Figure 5).

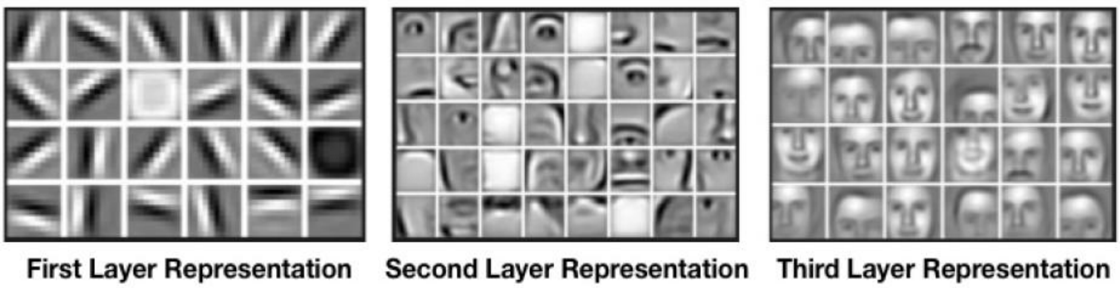


Figure 5. A visual representation of what images might look like on three CNN layers trained for image classification [30].

While CNNs are most commonly used for image recognition, they can also be applied to sound classification tasks, provided that the sound file has been transformed into an image before being used as input (e.g. MFCCs). Existing applications of CNNs include classifying cough sounds [31], lung sounds [32], heart sounds [33], or audio scene modelling [34].

Several small-scale studies have shown promising results in the field of lung sound analysis, but many improvements still need to be made to reach wider use [8]. A review done in 2017 notes that most researchers have simplified the task to only predict two or three classes (e.g. "healthy" / "not healthy"; "healthy" / "chronic disease" / "non-chronic disease") or to focus only on detecting abnormal lung sounds rather than respiratory diseases [32]. Studies done after 2017 seem to follow a similar trend [35–37]. Few studies were found that attempt at predicting six or more classes, such as this thesis. One notable exception is an experiment with 78 sound classes, achieving 62% test accuracy with a CNN and SVM model [32]. This suggests the novelty of this thesis while also hinting that predicting more than three classes is a challenging task that may require significant knowledge of the field and large amounts of data.

In addition to CNNs, recurrent neural networks (RNNs) and convolutional recurrent neural networks (CRNNs) have been used in sound signal analysis. RNNs have been used to classify cardiac arrhythmias (heart rhythm problems) [38] and to recognise speech [39–41]. CRNNs have been used for the classification of sound events (e.g. baby crying, gunshot) [42], for real-time speech enhancement [43,44] and for detecting multiple sounds and their directions from a single sound file [45,46]. While using RNNs and CRNNs is out of this thesis's scope, they would serve as an interesting comparison for the other models.

3.3 Dataset

The dataset used for experiments in this thesis was put together by researchers based in Greece and Portugal [47]. It contains 920 recordings of 126 patients' breathing. Each recording has information about each breathing cycle's start and end times and whether crackles or wheezes were present in each cycle.

In addition to recordings, the dataset contains the following data about patients: patient number, age, sex, BMI (for adults), weight (for children), height (for children), diagnosis. This data is stored in a tabular form.

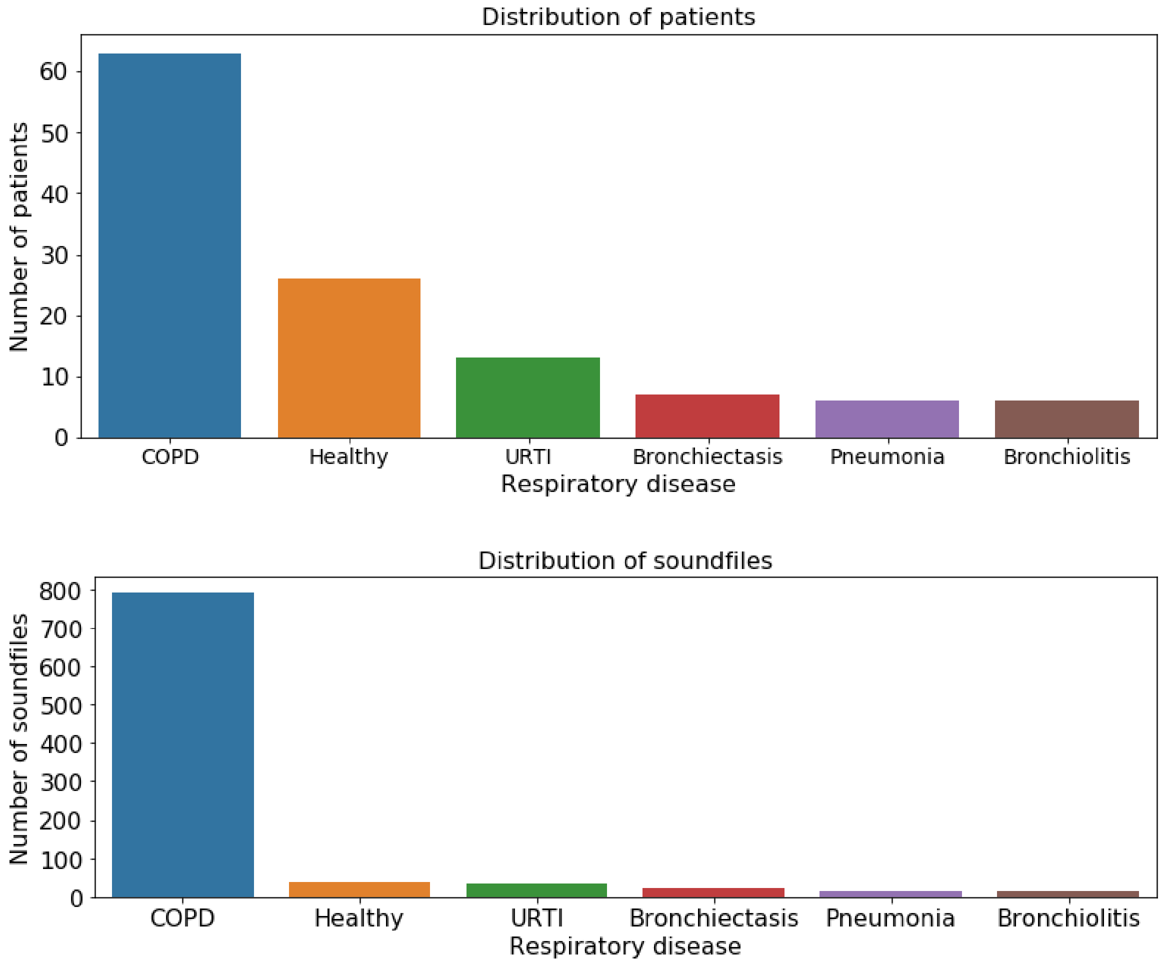


Figure 6. Database distribution of diseases across patients and sound files. The two least frequent classes (asthma and LRTI) are omitted from the figures because they were not used in this thesis.

Patient diagnosis can be one of 8 possible options: healthy, asthma, COPD, bronchiectasis, bronchiolitis, lower respiratory tract infection (LRTI), pneumonia, or upper respiratory tract infection (URTI). It is important to note that the distribution of diagnoses across patients and sound files is not uniform but rather heavily unbalanced, as shown in Figure 6. This makes it more challenging for models to correctly classify diagnoses.

3.4 Evaluation methods

The chosen metric for describing model performance is the macro-average F1-score. F1-score is the harmonic mean of precision and recall and was used because it combines both of those scores in a single metric. The macro-average is simply the arithmetic mean of F1-scores per each class. It was used because the dataset is highly unbalanced, and the macro-average considers each class equally. If a model only predicted the majority class correctly, the macro-averaged F1-score would be low.

Measures were taken to ensure that the deep learning results would be comparable to classical machine learning ones. Because deep learning models were trained on sound files and the rest on patient data with sound file statistics, the former would give diagnosis predictions for each sound, while the latter would give predictions for each patient. This made it more difficult to compare results. To fix this issue, the CNN model results were aggregated such that they would give predictions for each patient instead. This was achieved by looking at each patient's sounds, letting the CNN model predict each sound file's diagnosis, and then using the most frequent prediction as the output.

Because there were not enough patients to create one test set, which could accurately represent the entire dataset, cross-validation was used. Using this technique, a more reliable estimate of model performances can be captured even with a small dataset. Cross-validation works by splitting the data randomly into a training and testing set, training the model on the first one, testing on the other, and then repeating the process multiple times, afterwards combining the results. For each experiment, this was done in total 50 times to get the most reliable and robust results (i.e. 5-fold cross-validation was used ten times).

3.5 Frameworks

The two noteworthy libraries used in this thesis were Keras and Librosa. Keras is a high-level deep learning library. Its version 2.3.1 was used for training and evaluating the CNN models. Librosa is a python package made for audio and music analysis. Its version 0.7.2 was used for extracting sound features. Version 0.22.1 of a machine learning library Scikit-learn was used for building, training and evaluating the classical machine learning models. Numpy version 1.18.4 and pandas version 1.0.3 were used for data analysis, manipulation, and processing.

3.6 Computational resources

Initial work for this thesis was done using Google Colaboratory—a free online Python programming environment. However, this solution proved to be too slow for deep learning experiments. Thus, all experiments presented in this work were instead done in Jupyter notebooks that ran on the resources of the High Performance Computing Centre of the University of Tartu.

4 Experiments

In this chapter, the preprocessing and training details are provided for all experiments. This is split into two categories: classical machine learning and deep learning. A visual overview of the pipelines is provided in Figure 7.

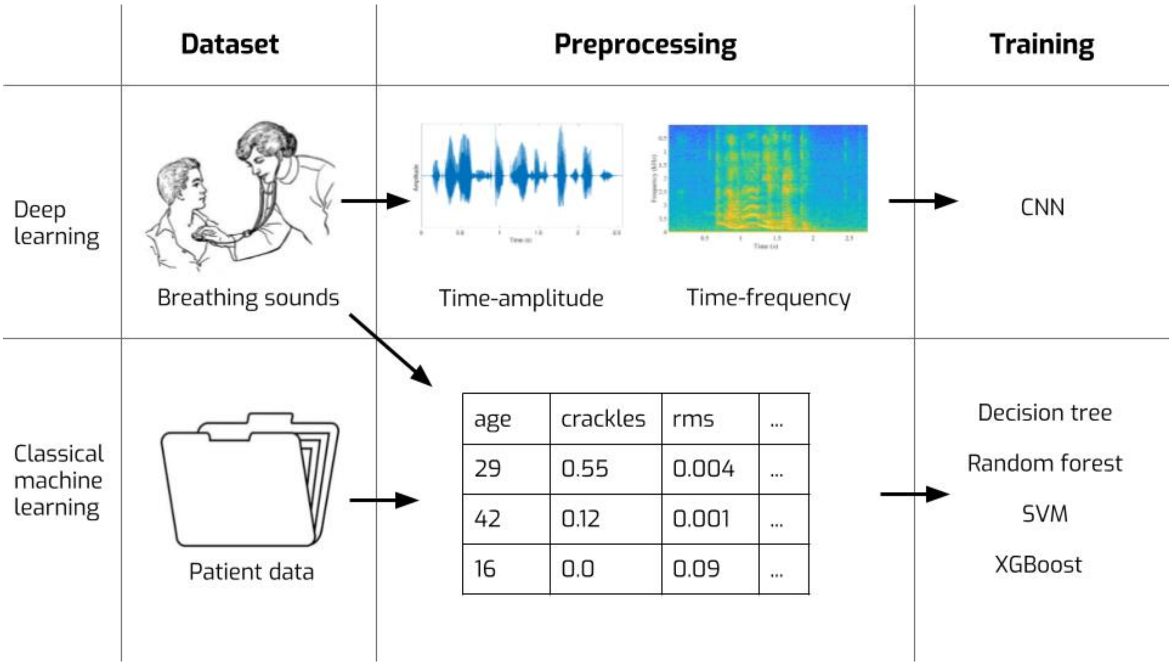


Figure 7. A high-level overview of the classical machine learning and deep learning pipelines.

In the classical machine learning pipeline, patient data and specific sound features were used as inputs to four different models. In the deep learning pipeline, only sound data was used. The sound files were converted into a visual representation of frequencies over time, which was then used as input for the deep learning model.

4.1 Classical machine learning pipeline

4.1.1 Preprocessing

Given the tabular patient data, the first task was discarding or replacing missing values. First, one patient was discarded, because they had missing sex and age values. Second, the BMI values for children had to be calculated. A few additional values were filled in using the values of patients with similar sex, age (+/- 5 years) and diagnosis. The two patients that still have some missing values were discarded.

In addition, since the distribution of diagnoses was far from uniform, the two diagnoses with the lowest frequency of occurence (2 LRTI patients and 1 asthma patient) were discarded. The height and weight columns were also discarded, since their information was already contained in the BMI column. In the end, 121 of the 126 patients remained, including data about their sex, BMI and diagnosis. This stage of preprocessing was inspired by a publicly available notebook on the same dataset [48].

The second task was extracting sound features. Firstly, using the recording annotations, information about the average number of crackles and wheezes per second was included. This intended to help differentiate between the healthy and the sick, as well as patients with pneumonia (which mainly causes crackles) and other diseases (which mainly cause wheezes).

Secondly, a set of 7 sound features (and their statistics) were chosen because they have been successfully used in sound classification tasks before [17,49]. The following features were extracted using the Librosa package and averaged over all sound files for each patient:

1. zero-crossing rate,
2. spectral centroid (mean, median, standard deviation (std)),
3. root mean square (mean, median, std),
4. spectral rolloff at 85% (mean, median, std),
5. spectral rolloff at 75% (mean, median, std),
6. spectral flatness (mean, median, std), and
7. spectral entropy.

The total number of features was 22, including 17 sound features, crackles and wheezes per second, BMI, sex, and age.

4.1.2 Training

The classical machine learning models used for training were the following: DT, RF, SVM, and XGBoost. The training was split into the following three experiments:

1. In **Experiment 1** ("Original model"), the models were run on all of the data. It would be expected that the models are prone to overfitting and thus mostly predict the majority class.
2. In **Experiment 2** ("Class weights"), class weights were introduced, making the models value each class equally during training. This was intended to make the models predict the minority classes more accurately.

3. In **Experiment 3** (“Fewer features”), the models were trained on seven features instead of 22. These features were chosen by using principal component analysis to view the data in 2-dimensions and by choosing the feature set in which classes were best separated. Principal component analysis is the process of decreasing the dimensionality of the data (decreasing the number of features), while preserving as much information as possible. The following seven features were chosen: "wheezes", "crackles", "age", "root_mean_square_mean", "spectral_entropy", "spectral_flatness_mean" and "zero_crossing_rate".

Because the test results fluctuated due to randomness, each experiment was conducted 10 times for each model. This increased the robustness of the results. Since each training run consisted of 5-fold cross-validation, there are in total 50 results for each experiment and for each model.

4.2 Deep learning pipeline

The deep learning pipeline was split into two phases: preprocessing and training. In the preprocessing phase, features were extracted from the original data and packed together into a format suitable for the CNN model. Three experiments were conducted, numbered 4 to 6.

4.2.1 Preprocessing

The preprocessing phase of the deep learning pipeline was as follows.

First, the LRTI and asthma patients’ sound files were discarded, as was done in the classical machine learning pipeline. This was because those classes contained too few patients and sound files. After that, each sound file was loaded into memory with a sample rate of 11 025 (samples per second). This was chosen after some experimentation showed that because sample rates above that did not increase results, but did considerably increase the data size. After that, each sound file was cut or randomly padded to be exactly 20 seconds long. 20 seconds was chosen because only a few sound files were longer, so minimal data was lost. After cutting or padding, every 512 samples of audio was converted into 40 MFCCs covering the frequency in a range of 50–2000 Hz. This range was chosen because respiratory sounds fall in this range and lower ranges would contain unnecessary information, such as heartbeats [50]. As a result, MFCCs of size 431x40 were extracted from each of 917 sound files.

For experiment number 6, an additional step was added to the preprocessing phase: data augmentation. Because the diagnosis distribution for sounds was even more unbalanced than for patients data augmentation was used to correct for this skew. The augmentation was done in the following way:

1. For COPD patients, MFCCs were extracted as normal.
2. For others, the sound file was split in pieces based on the respiratory cycles included in the annotations.
3. After that, the respiratory cycles were shuffled, creating a new sound file. MFCCs were extracted from that sound file as normal.
4. The number of new sound files created depended on the number of existing sounds such that the diagnosis distribution would be uniform in the end. For example, 20 new sounds had to be created for each pneumonia sound and 60 new sounds for each bronchiolitis sound.

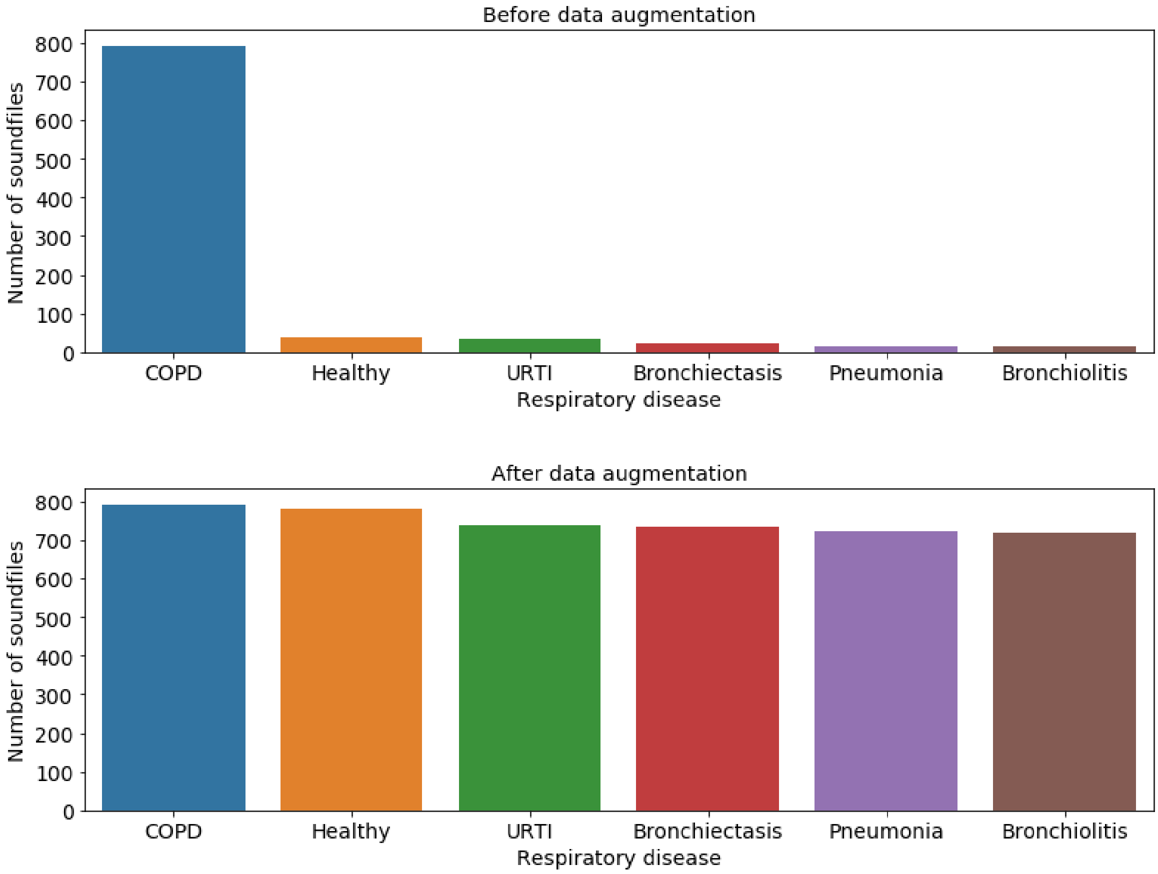


Figure 8. Distribution of diseases across sound files before and after augmentation.

In the end, 4491 sounds were created from the original 917 sounds, and the diagnosis distribution was much more uniform as a result (Figure 8).

4.2.2 Training

In the training phase, three experiments were conducted:

1. **Experiment 4** (“Original model”) was for baseline results.
2. In **Experiment 5** (“Class weights”), class weights were used to counter the unbalanced distribution of classes.
3. In **Experiment 6** (“Augmented data”), a method of data augmentation was used to generate additional sound files for the minority classes.

The CNN model used on all three experiments had three 2-D convolutional layers, fully-connected layers, pooling layers and dropout layers. The Adam optimiser was used for optimisation, and the loss function was categorical cross-entropy. Different optimisation algorithms were tried out, including RMSprop and Adagrad, but Adam seemed to give the best results. Different parameters were also tried out for the dropout layer, varying from 0.1 to 0.9, however, 0.4 seemed to be the best fit. A detailed summary of the model can be seen in Appendix 1.

All of the experiments were conducted with a batch size of 64, with 700 epochs and with a validation set which was 20% the size of the training set. The test set was roughly 20% of the entire dataset.

Just like with classical machine learning, 5-fold cross-validation runs were performed ten times in each experiment to increase robustness, resulting in 50 individual results.

5 Results

In this chapter, the experiment results are presented. The findings are explained, analysed, and the difficulties of predicting certain diseases on this particular dataset are brought out.

5.1 Classical machine learning results

A comparison of the macro-averaged F1-scores on test sets for classical ML algorithms is presented in Table 1, which shows the average and the standard deviation of 50 test results.

Table 1. Classical machine learning test results. The best result for every model is indicated in bold.

	SVM	**DT**	**XGBoost**	**RF**
Original model	0.4856 +/- 0.12	0.4869 +/- 0.1	0.4801 +/- 0.12	0.3597 +/- 0.08
Class weights	0.4918 +/- 0.11	0.5277 +/- 0.11	0.4571 +/- 0.11	0.4052 +/- 0.1
Fewer features	**0.6189 +/- 0.13**	**0.5410 +/- 0.1**	**0.5091 +/- 0.13**	**0.4644 +/- 0.13**

Several notable conclusions can be made from these results. Firstly, SVM seems to perform the best. The performance difference between the best SVM model and the next best model (DT with fewer features) is 0.0779 (7.8%). This is not too surprising since the model has performed successfully in similar experiments before [8]. However, some aspects of the experiments might have artificially inflated this result. This needs to be pointed out for transparency. One possible reason for the high SVM scores is that during preprocessing, some information from the validation sets of each cross-validation run may have ended up in the training sets due to using normalisation, parameter tuning, and feature selection on the entire dataset. Even so, the results suggest SVM might have promise in future applications.

A surprising find is that RF and XGBoost have, on average, inferior performance compared to relatively simpler models (DT and SVM). They might have poor results because very few features were actually required to predict well. Since they select a random subset of features for each sub-tree, some trees might have been left with only poor features, which brought the overall performance down. Evidence for this is apparent when looking at feature importances for random forest and decision tree models (Figure 9). DT gives a lot of importance to ‘age’, whereas RF uses other features more often.

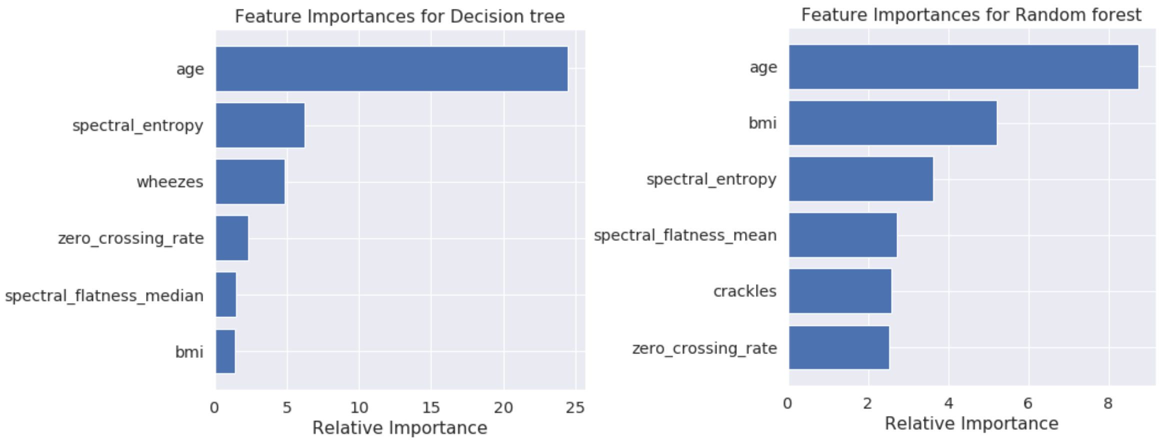


Figure 9. Feature importance for the decision tree and random forest. The six most important features are presented.

In the case of all these results, it is important to note that the macro-averaged F1-score is used, which weighs all classes equally. In fact, when each sample is weighted equally, the results are much different and less unexpected: in that case, XGBoost and RF do get better scores than DT (Table 2).

Table 2. Best micro- and macro-averaged results for classical machine learning models.

	SVM	**DT**	**XGBoost**	**RF**
Best micro-averaged F1-score	**0.8050 +/- 0.06**	0.7166 +/- 0.06	0.7685 +/- 0.06	0.7653 +/- 0.04
Best macro-averaged F1-score	**0.6189 +/- 0.13**	0.5410 +/- 0.1	0.5091 +/- 0.13	0.4644 +/- 0.13

No matter which metric to look at, DT, RF, and SVM models all improved with class weights, as is expected. Adding class weights increased the importance of minority classes for the models, which in turn improved the results. The only exception here is XGBoost, which surprisingly decreased with class weights, though by fairly little.

Also evident from the results is the fact that all models improved after decreasing the number of features. This indicates that the initial set of features contained much unnecessary information, while the second set of features was chosen more wisely.

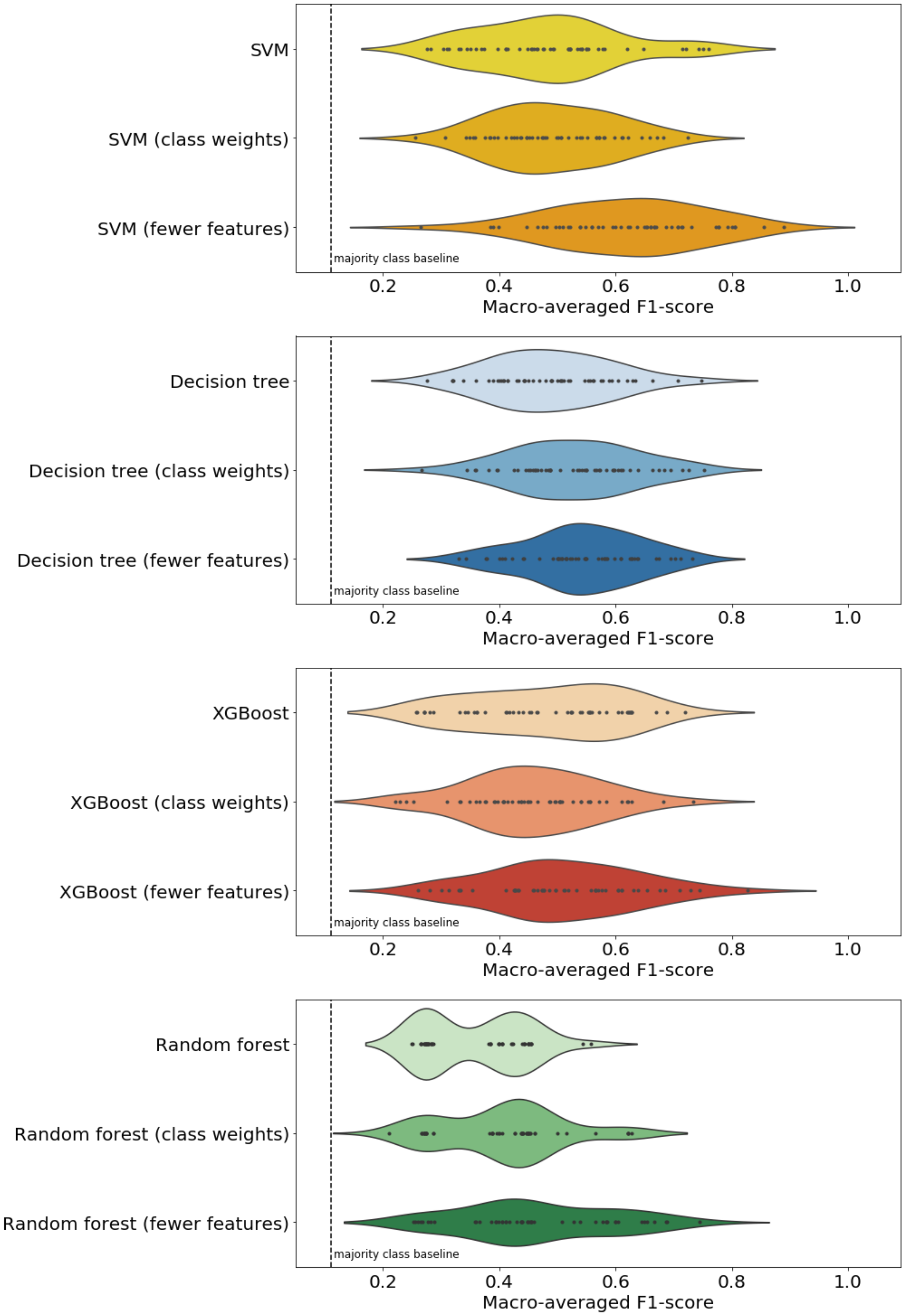


Figure 10. Classical machine learning test results. Each violin plot shows 50 results for each model. The "majority class baseline" at 0.1 indicates results from just predicting "COPD".

Finally, it is important to note that the difference between the three best performing models (SVM, DT, XGBoost) all fall within a single standard deviation range. In fact, the results had very large standard deviations in general: each individual cross-validation run gave varying results even with the same model and dataset. This can be visually seen in Figure 10. These wide fluctuations in results suggest that more data would be required to give conclusive results.

5.2 Deep learning results

A comparison of the macro-averaged F1-scores on test sets for the convolutional neural network models is presented in Table 3, together with the best and worst-performing classical ML methods. The table shows the average and the standard deviation of 50 individual results.

Table 3. Deep learning test results compared with SVM and RF test results. The best result for each method is indicated in bold. "N/A" indicates that such an experiment was not performed on that model.

	CNN	**SVM**	**RF**
Original model	0.2416 +/- 0.09	0.4856 +/- 0.12	0.3597 +/- 0.08
Class weights	**0.3041 +/- 0.11**	0.4918 +/- 0.11	0.4052 +/- 0.10
Augmented data	0.2120 +/- 0.08	N/A	N/A
Fewer features	N/A	**0.6189 +/- 0.13**	**0.4644 +/- 0.13**

Since the number of data points for each class (other than COPD) was less than 100, it was expected that deep neural networks would struggle with classification. The experiment results confirmed this intuition—even the worst-performing ML model achieved 52% better results than the best CNN model.

There are several possible causes for such poor results. One clear factor was that the dataset used for deep learning (sound files) was even more unbalanced than the one used for machine learning (patient data with sound features). The difference between the first two most frequent classes was more than 10-fold (almost 800 for COPD and less than 80 for healthy patients). This made it very challenging to train the model, as it mostly just predicted COPD for the diagnosis. This is evident from a sample confusion matrix in Figure 11. It seems that there just was not enough data points for the minority classes.

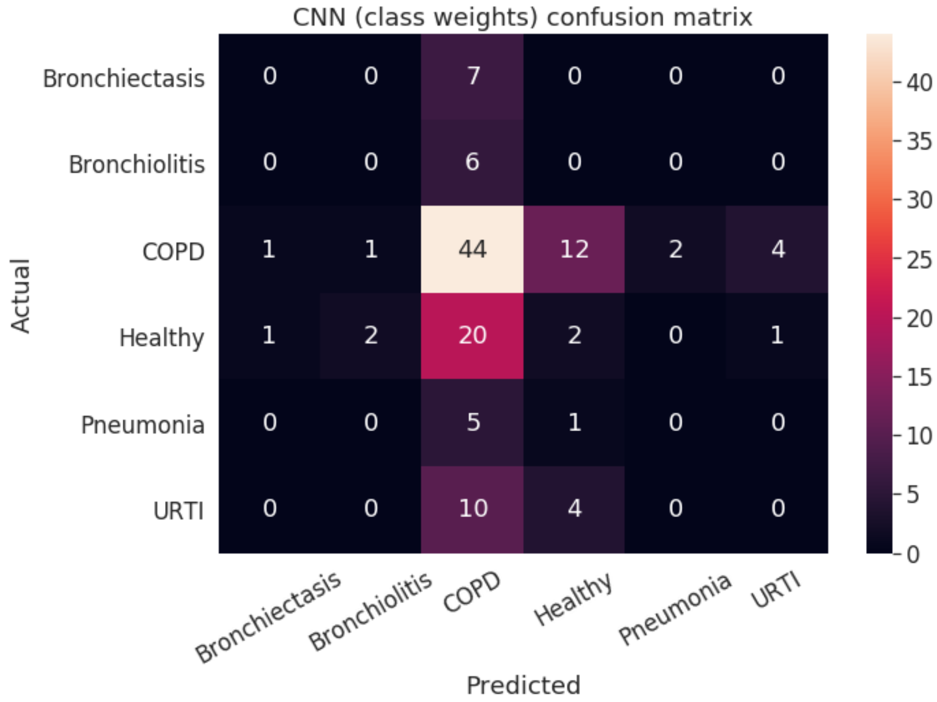


Figure 11. Confusion matrix for a CNN model of a single cross-validation fold.

Two measures against this unbalancedness were taken: class weights and augmented data. Predictably, class weights did improve the results considerably. However, they still fell short of any classical ML results.

But, the results from augmented data were surprising. They were in fact worse than the original CNN model with no class weights or augmentation. Several key factors may have influenced this. First, inappropriate augmentation may have harmed the performance. It may be possible that the attempted augmentation algorithm was adding noise to the data instead of additional data. The algorithm shuffled breathing cycles around to create new breathing sounds. This particular augmentation may not be acceptable for this kind of data. The precise sequence of breathing cycles might have contained useful information that was lost during shuffling and made it more difficult to predict diagnoses. Another possible reason is that the breathing cycles were annotated incorrectly and as a result, after being reshuffled, this procedure produced unnatural and therefore fruitless recordings.

There are several alternative approaches to the augmentation employed in this thesis:

1. a portion of COPD could be discarded,
2. minority class sound files could be duplicated,
3. the original sound could be amplified and
4. pitch and speed can be changed.

It is plausible that some combination of these alternative approaches could give better results, however, this seems unlikely that they would exceed the results from classical machine learning models on this particular dataset.

Another problem which made classification more tricky was the high level of data noise. Some sound files had loud irrelevant noises on the background (e.g. human speech and ticking clocks). Measures were taken to reduce this (for example, limiting the frequency range to 50-2000 Hz), but it may not have been enough. To improve the results, it would be best to consider more advanced algorithms to extract only the important breathing sounds from the data. This could be done by training another algorithm to recognise common abnormal breathing sounds (crackles and wheezes) and to discard all other information from the sound.

Finally, it might be unfair to directly compare ML models with CNN because the former had more data: information about each patient, such as their age and BMI. In fact, the decision tree made a significant amount of decisions just based on the patient's age. Perhaps if the ML models had been trained on just sound data, the deep learning model would have come out ahead.

5.3 Challenges of disease prediction

Most of the difficulties in classifying diseases on this dataset become apparent when looking at the principal component analysis (PCA) plot (Figure 12).

It seems that by using sound features, it easy to differentiate between COPD and healthy patients, as healthy people should have no crackles or wheezes, but COPD patients do. It also seems to be easy to tell apart bronchiolitis patients from others, as bronchiolitis is mostly only diagnosed for patients under the age of two, and the tabular data contains the age of the patient. In fact, "age" was by far the most important feature for the decision tree.

However, it seems to be much more difficult to differentiate between bronchiectasis, pneumonia and COPD patients. As bronchiectasis can be the result of COPD, the similarities make sense. Confusing pneumonia with other diseases makes less sense because pneumonia should mainly cause crackles, while others should mainly cause wheezes.

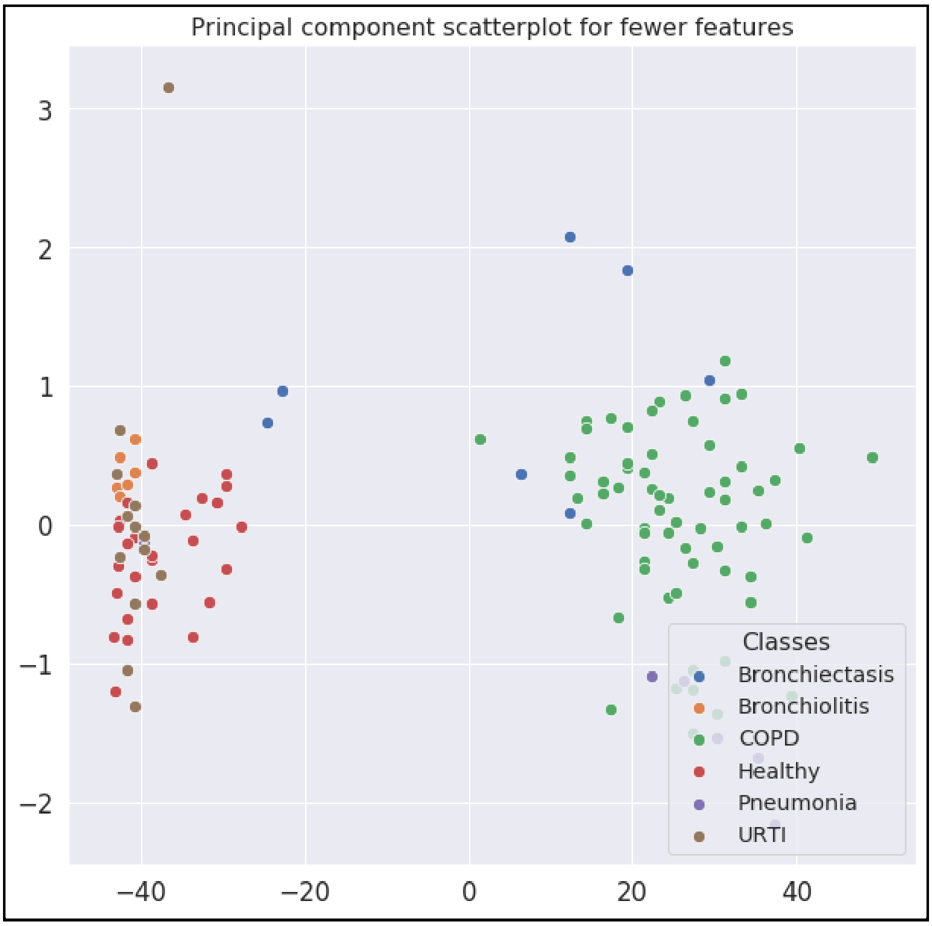


Figure 12. Scatterplot for two principal components on data with seven sound features.

URTI was by far the most difficult to classify, naturally because the lungs of URTI patients and healthy patients should sound identical. The infection for URTI patients is not in the lungs but in the upper respiratory tract (nose, sinuses), as the name would suggest.

Therefore, the easiest improvements could likely be made in differentiating pneumonia patients from others. This could be done by taking into account the location of the sounds in the lung (as pneumonia mostly affects alveoli). When recording new sounds, including the physical position of the patient during breathing might help too. Different patient positions can play a role in how well lung sounds can be heard [12].

6 Conclusion

In this chapter, a summary of the results is presented, and a list of possible improvements upon this work is listed.

6.1 Results summary

The purpose of this thesis was to create a codebase for predicting respiratory diseases using lung sounds. This is important because these diseases (including COVID-19) are one of the leading causes of death globally. Improved accessibility and accuracy of diagnoses could significantly decrease the global disease burden. This codebase would help future researchers get started with medical machine learning in the future with less effort and less time.

A comprehensive comparison between classical machine learning and deep learning on a dataset of lung sounds was presented. In total, six experiments were carried out with five methods: four classical machine learning methods (DT, RF, SVM and XGBoost) and one deep learning method (CNN).

The test results showed that simple classical machine learning methods, like support vector machine and decision tree, performed best when looking at all prediction classes equally (i.e. using the macro-averaged F1-score as the primary metric). Using class weights and decreasing the number of features improved results. The best scores overall were achieved by SVM and decision tree models with decreased features (0.62 and 0.54). The convolutional neural network models achieved very poor results compared to others (best score 0.3). A novel data augmentation method was used to improve performance, but it failed to do so.

Various poor features of the dataset made accurate predicting more challenging: it is highly unbalanced, contains a small number of patient data points, and the sound data is often noisy. Because of these reasons, it can be concluded that more and better quality data should be added before making substantial claims on the overall usefulness of the models.

All the goals of this thesis were achieved, and the source code is publicly available in a Github repository [1].

6.2 Future work

Here are a few ideas on how to improve upon this work:

1. Using information about the patient’s position during breathing or the location of the stethoscope on the patient’s chest might help distinguish certain diseases from others (such as pneumonia).
2. Additional methods for visually representing sounds might work better than MFCCs. Options include discrete wavelet transform (DWT), chromatograms, scalograms or spectrograms.
3. Alternatives to convolutional neural networks might achieve better results. Recurrent neural networks (RNN) or convolutional recurrent neural networks (CRNN) could be used.
4. URTI patients should be excluded from the dataset because their disease produces no lung sounds and they are indistinguishable from healthy people.
5. Using more advanced algorithms for noise reduction or for extracting abnormal breathing sounds from the data is suggested.
6. Alternative methods for data augmentation might produce better results.

However, the easiest improvement would be to find a larger and more balanced database or attempt at classifying something more simple, such as the healthy and the sick or crackles and wheezes.

7 References

[1] Annilo R. Github Repository. https://github.com/rannilo/respiratory-disease-prediction (04.01.2021)

[2] Forum of International Respiratory Societies, European Respiratory Society. The global impact of respiratory disease. 2017.

[3] Fred H.L. Drawbacks and limitations of computed tomography: views from a medical educator. Texas Heart Institute Journal, 2004, Volume 31, Issue 4, pp 345–348.

[4] Wilkinson I.B., Raine T., Wiles K. et al. Oxford Handbook of Clinical Medicine. 10th ed. Vol. 1. Oxford University Press. 2017.

[5] Safdar S., Zafar S., Zafar N., Khan N.F. Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Artificial Intelligence Review, 2018, Volume 50, Issue 4, pp 597–623.

[6] Thomsen K., Iversen L., Titlestad T.L., Winther O. Systematic review of machine learning for diagnosis and prognosis in dermatology. Journal of Dermatological Treatment, 2020, Volume 31, Issue 5, pp 496–510.

[7] Hwang E.J., Park S., Jin K.-N. et al. Development and Validation of a Deep Learning–based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs. Clinical Infectious Diseases, 2019, Volume 69, Issue 5, pp 739–747.

[8] Das N., Topalovic M., Janssens W. Artificial intelligence in diagnosis of obstructive lung disease: current status and future potential. Current Opinion in Pulmonary Medicine, 2018, Volume 24, Issue 2, pp 117–123.

[9] UNSHAW. Basic diagram of the Human Respiratory System. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Human_Lungs.gif (19.12.2020)

[10] Moussavi Z. Fundamentals of Respiratory Sounds and Analysis. 1st ed. Morgan & Claypool. 2006.

[11] Rogers K. The Respiratory System. 1st ed. Britannica Educational Publishing. 2011.

[12] Wipf J.E., Lipsky B.A., Hirschmann J.V. et al. Diagnosing Pneumonia by Physical Examination: Relevant or Relic? Archives of Internal Medicine, 1999, Volume 159, Issue 10, pp 1082.

[13] McLuckie A. Respiratory Disease and its Management. Springer.

[14] Lerch A. An Introduction to Audio Content Analysis. John Wiley & Sons, Inc. 2012.

[15] Ajibola Alim S., Khair Alang Rashid N. Some Commonly Used Speech Feature Extraction Algorithms. IntechOpen. 2018.

[16] Tian Y., Zhang H., Xu W. et al. Spectral Entropy Can Predict Changes of Working Memory Performance Reduced by Short-Time Training in the Delayed-Match-to-Sample Task. Frontiers in Human Neuroscience, 2017, Volume 11, pp 437.

[17] Kos M., Kačič Z., Vlaj D. Acoustic classification and segmentation using modified spectral roll-off and variance-based features. Digital Signal Processing, 2013, Volume 23, Issue 2, pp 659–674.

[18] Maningo E. Understanding what does RMS stands for in Audio: Definition & Details. AudioRecording. https://www.audiorecording.me/understanding-what-does-rms-stands-for-in-audio-definition-details.html (23.04.2020)

[19] Madhu N. Note on measures for spectral flatness. Electronics Letters, 2009, Volume 45, Issue 23, pp 1195.

[20] Přibil J., Přibilová A. Spectral Flatness Analysis for Emotional Speech Synthesis and Transformation. Berlin, Heidelberg: Springer Berlin Heidelberg. 2009. (Lecture Notes in Computer Science; vol. 5641).

[21] Flach P.A. Machine learning: the art and science of algorithms that make sense of data. Cambridge ; New York: Cambridge University Press. 2012.

[22] Song Y.-Y., Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry, 2015, Volume 27, Issue 2, pp 130–135.

[23] Fallon B., Ma J., Allan K. et al. Opportunities for prevention and intervention with young children: lessons from the Canadian incidence study of reported child abuse and neglect. Child and Adolescent Psychiatry and Mental Health, 2013, Volume 7, Issue 1, pp 4.

[24] Fristrup K.M., Watkins W.A. Marine Animal Sound Classification. 1993. https://apps.dtic.mil/dtic/tr/fulltext/u2/a289294.pdf (28.02.2020)

[25] Pavlopoulos S.A., Stasis A.C., Loukis E.N. A decision tree - based method for the differential diagnosis of Aortic Stenosis from Mitral Regurgitation using heart sounds. BioMedical Engineering OnLine, 2004, Volume 3, Issue 1, pp 21.

[26] Tin Kam Ho. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, Volume 1, pp 278–282.

[27] Hastie T., Tibshirani R., Friedman J.H. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, NY: Springer. 2009. (Springer series in statistics).

[28] Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, 2016.

[29] Github. https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions (21.12.2020)

[30] Nielsen C. Medium. https://medium.com/chris-nielsen/making-my-own-deep-learning-image-recogniser-d707435a788e (21.12.2020)

[31] Amoh J., Odame K. Deep Neural Networks for Identifying Cough Sounds. IEEE Transactions on Biomedical Circuits and Systems, 2016, Volume 10, Issue 5, pp 1003–1011.

[32] Aykanat M., Kılıç Ö., Kurt B., Saryal S. Classification of lung sounds using convolutional neural networks. EURASIP Journal on Image and Video Processing, 2017, Volume 2017, Issue 1, pp 65.

[33] Singh S.A., Majumder S., Mishra M. Classification of short unsegmented heart sound based on deep learning. 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 2019.

[34] Chen H., Zhang P., Bai H. et al. Deep Convolutional Neural Network with Scalogram for Audio Scene Modeling. Interspeech 2018, 2018.

[35] Pham L., McLoughlin I., Phan H. et al. Robust Deep Learning Framework For Predicting Respiratory Anomalies and Diseases. arXiv:200203894 [cs, eess], 2020. http://arxiv.org/abs/2002.03894 (21.12.2020)

[36] Jayalakshmy S., Sudha G.F. Scalogram based prediction model for respiratory disorders using optimized convolutional neural networks. Artificial Intelligence in Medicine, 2020, Volume 103, pp 101809.

[37] Perna D., Tagarelli A. Deep Auscultation: Predicting Respiratory Anomalies and Diseases via Recurrent Neural Networks. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Cordoba, Spain: IEEE. 2019., p. 50–55.

[38] Latif S., Usman M., Rana R., Qadir J. Phonocardiographic Sensing using Deep Learning for Abnormal Heartbeat Detection. arXiv:180108322 [cs], 2018. http://arxiv.org/abs/1801.08322 (17.02.2020)

[39] Li J., Mohamed A., Zweig G., Gong Y. LSTM time and frequency recurrence for automatic speech recognition. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015.

[40] Sainath T.N., Li B. Modeling Time-Frequency Patterns with LSTM vs. Convolutional Architectures for LVCSR Tasks. 2016.

[41] Li J., Mohamed A., Zweig G., Gong Y. Exploring multidimensional lstms for large vocabulary ASR. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai: IEEE. 2016., p. 4940–4944.

[42] Kao C.-C., Wang W., Sun M., Wang C. R-CRNN: Region-based Convolutional Recurrent Neural Network for Audio Event Detection. arXiv:180806627 [cs, eess], 2018. http://arxiv.org/abs/1808.06627 (22.02.2020)

[43] Tan K., Wang D. A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement. Interspeech 2018, 2018.

[44] Zhao H., Zarar S., Tashev I., Lee C.-H. Convolutional-Recurrent Neural Networks for Speech Enhancement. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[45] Adavanne S., Politis A., Nikunen J., Virtanen T. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. IEEE Journal of Selected Topics in Signal Processing, 2018, Volume 13, Issue 1, pp 34–48.

[46] Pratik P., Jee W.J., Nagisetty S., Mars R., Lim C. Sound Event Localization and Detection using CRNN Architecture with Mixup for Model Generalization. Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 2019.

[47] Rocha B.M., Filos D., Mendes L. et al. A Respiratory Sound Database for the Development of Automated Classification. In: Precision Medicine Powered by pHealth and Connected Health, Singapore: Springer Singapore. 2018., p. 33–37. (IFMBE Proceedings; vol. 66).

[48] gizemt, İymen G., Saif M. Multi-class Classification by Decision Tree. https://www.kaggle.com/gizemtanriver/multi-class-classification-by-decision-tree (15.05.2020)

[49] Tang H., Dai Z., Jiang Y., Li T., Liu C. PCG Classification Using Multidomain Features and SVM Classifier. BioMed Research International, 2018, Volume 2018, pp 1–14.

[50] Andrès E., Gass R., Charloux A., Brandt C., Hentzler A. Respiratory sound analysis in the era of evidence-based medicine and the world of medicine 2.0. Journal of Medicine and Life, 2018, Volume 11, Issue 2, pp 89–106.

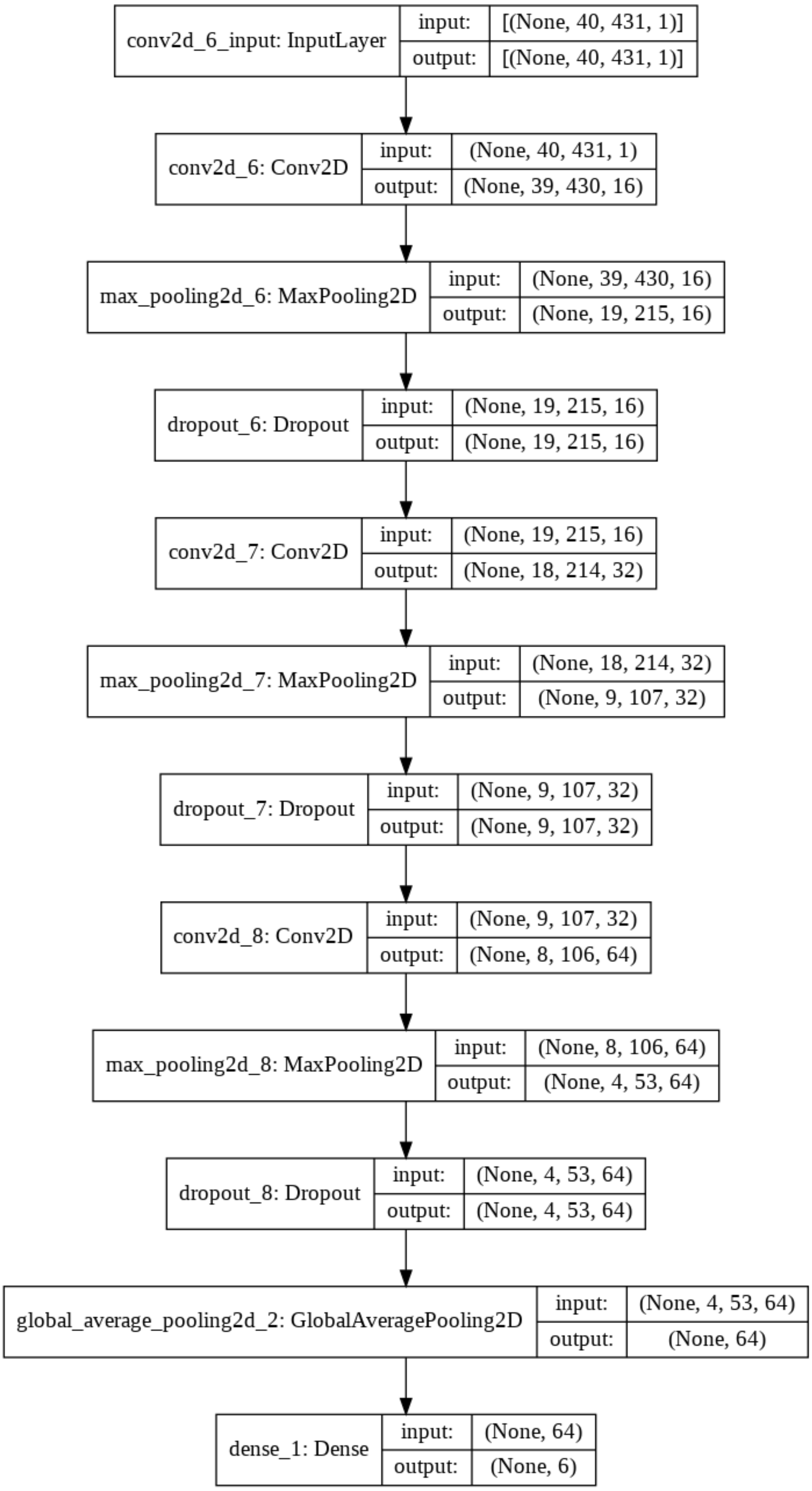
Acknowledgements

This work was carried out in the High Performance Computing Center of University of Tartu.

Many thanks to Dmytro Fishman for his continued support throughout the process of completing this thesis: from suggesting online courses and explaining difficult machine learning concepts to giving his final notes on the writing.

Appendices

I. Deep learning model summary



II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Richard Annilo** (date of birth 22.08.1998),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:
    a. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and
    b. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

    „**Predicting Respiratory Diseases from Lung Sounds Using Machine Learning**", supervised by Dmytro Fishman

2. am aware of the fact that the author remains these rights.
3. certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Richard Annilo

Tartu, 12.01.2021