2022

2D respiratory sound analysis to detect lung abnormalities

Rafia Sharmin Alice

2D RESPIRATORY SOUND ANALYSIS TO DETECT LUNG ABNORMALITIES

By

Rafia Sharmin Alice
B.Sc., Khulna University, Bangladesh, 2019

A Thesis Submitted in Partial Fulfillment of
The Requirements for the Degree of Master of Science

Department of Computer Science

Master of Science Program
In the Graduate School
The University of South Dakota
May 2023

The members of the Committee appointed to examine
the Thesis of Rafia Sharmin Alice
find it satisfactory and recommend that it be accepted.

DocuSigned by:
K C Santosh
BBEF270EA578492...

Chairperson

DocuSigned by:
Doug Goodman
3C0E646BD1404F3...

DocuSigned by:
Pere Miro
9C2DB1B5F18A4BB...

Abstract

In this paper, we analyze deep visual features from 2D data representation(s) of the respiratory sound to detect evidence of lung abnormalities. The primary motivation behind this is that visualcues are more important in decision-making than raw data (lung sound). Early detection and prompt treatments are essential for any future possible respiratory disorders, and respiratory sound is proven to be one of the biomarkers. In contrast to state-of-the-art approaches, we aim atunderstanding/analyzing visual features using our Convolutional Neural Networks (CNN) tailored Deep Learning Models, where we consider all possible 2D data such as Spectrogram, Mel-frequency Cepstral Coefficients (MFCC), spectral centroid, and spectral roll-off. In our experiments, using the publicly available respiratory sound database named ICBHI 2017 (5.5 hours of recordings containing 6898 respiratory cycles from 126 subjects), we received the highest performance with the area under the curve of 0.79 from Spectrogram as opposed to 0.48 AUC from the raw data from a pre-trained deep learning model: VGG16. We also used machine learning algorithms using reliable data to improve Our study proved that 2D data representation could help better understand/analyze lung abnormalities as compared to 1D data. Our findings are also contrasted with those of earlier studies. For purposes of generality, we used the MFCC ofneutrinos to determine if picture data or raw data produced superior results.

Thesis Advisor: DocuSigned by: KC Santosh BBEF270EA578492...

KC Santosh, Ph.D.

Acknowledgments

I would like to express my special gratitude and thanks to my thesis advisor Dr. KC San- tosh for his constant instruction, guidance, and encouragement throughout my thesis completion. Without his thoughtful encouragement and careful supervision, this thesis would never have taken shape. I also thank my thesis committee members, Dr. Doug Goodman, and Dr. Pere Miro for their contributions to the direction and richness of this work. I am also grateful for the dataset (ICBHI17). I would like to thank Nick Rasmussen to help me understand the concept of sound processing. I would like to thank Siva Allu for making the extension of this thesis possible by providing the dataset.

Dedication

To all the people who believed in me in times even when I didn't believe in myself.

Table of Contents

Chapter 1 Introduction

Summary: The root cause of death and disability worldwide are respiratory disorders. The highest disease load was seen in the world's most underdeveloped areas. According to experts, aging and risk factors like smoking, air pollution, and body weight are also important. A significant public health issue, chronic obstructive pulmonary disease affects about 65 million people and is the third biggest cause of mortality globally, accounting for 3.91 million fatalities in 2017, according to estimates. Chronic respiratory disease-related fatalities increased by 18% between 1990 and 2017, from 3.32 million in 1990 to 3.91 million in 2017 [1]. Since early detection and prompt treatment are essential for respiratory disorders, the audio of the respiratory sounds has been proven to be highly helpful. In this chapter, we present the motivation to attempt to improve respiratory disease detection using image data generated from respiratory audio sound instead of raw data and to see if image data gives improved results from the audio sound. After that, we will explain the problem-solving way to reach the goal and our contribution.

Key topics: Motivation, goal, and contribution.

Organization: The chapter is structured as follows. In section 1.1, we will give a detailed explanation of the context and problem of our thesis work, and goal of our work section, and what we are trying to achieve here. After that, section 1.2 contains the motivation for our work. Section 1.3 contains our contribution regarding our proposed work. Finally, section 1.4 describes "what's next," in chapter 2

1.1 Context, Problem, and Goal

Respiratory diseases are hard to treat if not detect early. Various respiratory diseases are the leading cause of death every year. Such as asthma, the most prevalent chronic childhood illness impacting 14% of all children worldwide, affects around 334 million individuals. Pneumonia, one of the most common causes of death in children under the age of five, kills millions of people every year. The most prevalent fatal infectious illness, TB affects about 10 million people and kills 1.4 million each year. The worst malignancy, lung cancer claims 1.6 million lives each year. A chronic respiratory condition causes the early deaths of 4 million people worldwide. Five of the top 30 causes of death are respiratory diseases: COPD is third, followed by lower respiratory tract infections, tracheal, bronchial, and lung cancer, tuberculosis (TB), and asthma [1]. Acute or chronic respiratory diseases affect more than 1 billion people worldwide. The unfortunate truth is that chronic respiratory diseases cause 4 million premature deaths annually [2]. Young children and infants are especially vulnerable. Pneumonia is the world's biggest cause of death for children under 5 years old, accounting for a total of 9 million deaths per year
[1].

People sometimes take their ability to breathe and the state of their respiratory system for granted, yet the lung is a crucial organ that is susceptible to airborne illness and damage. Because of the fragility of the lung, the survival rates for lung transplant patients are not as good as for other solid organ transplants, with a five-year survival rate of about 50-60%. The biggest limiting factor in lung transplants is having enough suitable lung donors. Diseases of the respiratory system have a substantial impact on people's social, economic, and healthy lives. The most significant factor influencing death and disability rates was social deprivation, with the highest rates observed in the world's most impoverished areas. More developed nations had lower mortality rates due to better access to medical care and more effective treatments.

Assistive solutions to problems in the medical arena have been made possible thanks in large part to technologies like machine learning and deep learning. Medical imaging informatics increase the predictive accuracy of early and timely disease identification. Due to the lack of skilled human resources, medical professionals are grateful for such technical support because it enables them to handle more patients. Aside from serious illnesses like cancer and diabetes, the prevalence of respiratory conditions is gradually increasing and posing a life-threatening threat to society.

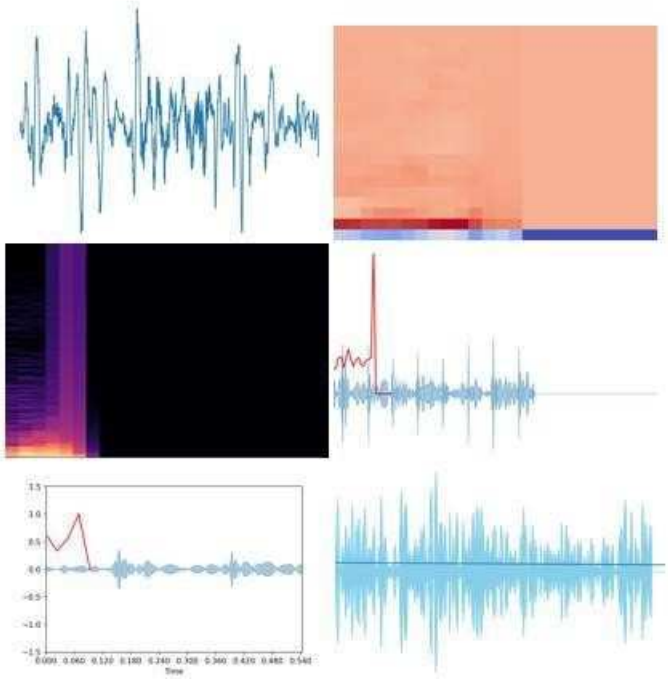

Fig. 1. Qualitative output: Raw data (upper-left), MFCC (upper-right), Spectrogram (middle left), Spectral centroid (middle-right), Spectral roll-off (lower-left), and zero crossing rate (lower-right)

In the seminal work by Zeiler and Fergus on deep learning visualization [9], a method known as deconvolutional networks allowed projection from a model's learned feature space back to the pixel space, or, to put it another way, gave us a glimpse at what neural networks were seeing in large sets of images. Their method and findings provide a debugging tool for developing a model and shed light on the kinds

of properties deep neural networks are learning at particular layers in the model. Following this sort of work we are attempting to understand if visualization can improve the detection of lung abnormality from audio data. As the picture says thousands of words, image data should provide better insight than raw data due to having a large number of feature sets. Our goal is to process the sound data and apply it to the different deep-learning models to compare the results with raw data.

1.2 Motivation

There are numerous studies on lung and heart sound identification because lung and heart diseases continue to be the primary causes of death in the world [4]. Since then, there have been numerous advancements in the processing of audio recordings made in noisy settings. Furthermore, innovative techniques like machine learning and deep learning significantly advance the field. These methods make significant contributions to computer vision and audio analysis. This helps extract more pertinent information from the respiratory sounds and speeds up diagnosis, improving treatment effectiveness [5]. Therefore, a computer program designed to detect irregularities in respiratory sounds may be quite useful in clinical diagnosis. Many scientists are investigating how to combine speech and signal processing tools with image analysis-based tools so that clinicians can predict or guess the existence of respiratory disorders based on verbal communication before they even begin the X-ray screening or other procedures [6].

The simple and non-invasive auscultation process can be used to get respiratory sounds. After utilizing a stethoscope to listen for lung disease, doctors can effectively evaluate and diagnose the condition using auscultation. Since there is no need for inside bodily involvement, this procedure is cheap and simple. On the other hand, traditional stethoscopes may be susceptible to external noise sounds, cannot filter the body's audio frequencies during auscultation, and cannot produce long- term recordings for tracking the progression of the disease. Due to the worldwide increase in respiratory disorders, medical research has become more interested in incorporating potential audio signal analysis-based techniques.

Since a few decades ago, computer science has been steadily advancing our ability to automatically analyze media data, and with the use of diagnosis tools, we are able to interpret image and audio data. Therefore, computer science could aid nursing staff or doctors in diagnosis by suggesting quicker and more reliable instruments and by providing the patient with adaptable tools for medical monitoring. Therefore we are trying to learn if image data are more effective to improve the accuracy of the detection of abnormalities.

1.3 Contribution

In the machine learning and deep learning domain, understanding how data is used is crucial since it facilitates its analysis. Visualization is an effective technique for understanding and improving deep learning models. The use of visual encodings to convert abstract data into useful representations allows data visualization and visual analytics to effectively convey information and uncover insights. Deep learning is a particular set of techniques that, when given a dataset, learns what features are relevant to the task [7]. As an example, we have demonstrated different data formats (in different spaces) in Fig 1: raw data, MFCC, Spectrogram, Spectral centroid, Spectral roll- off, and zero-crossing rate. This contrasts with

typical machine learning approaches, which employ a dataset with known features such as a collection of autos where the collection of automobiles represents a dataset, with known makes, models, and colors, these represent features. This is helpful for datasets that do not explicitly have tabular elements, such as a collection of pictures, a collection of texts, or an audio library. Deep artificial neural networks are the preferred model architecture for these kinds of models. As complicated deep learning models are difficult to train and comprehend, interactive interfaces and visualizations have been built and developed to assist individuals in understanding what models have learned and how they make predictions. A crucial step in studying and discovering relationships between various entities is feature extraction [8]. To transform the provided audio data into a format that the models can understand, feature extraction is used. It is a method that provides a clear explanation for the majority of the data. For classification, prediction, and recommendation algorithms, feature extraction is necessary.

This study is frequently credited with popularizing visualization in the computer vision and deep learning areas in recent years by demonstrating visualization as a potent tool for understanding and enhancing deep models. Here

1. We have segmented each audio recording as an audio clip containing high deviations across its entire length, its analysis is not trivial. Therefore, each audio clip is broken down into smaller segments to facilitate analysis.

2. We have generated different sorts of image data. In total, we have created six sorts of image data.

3. We have applied four types of deep learning models and two types of machine learning models to each of the image data in order to find out if image data gives a better result than raw data.

4. For generalization we have applied Vgg16 on the MFCC of neutrino data and Machine learning models on raw data and compared the results.

1.4 Thesis outline

The rest of our thesis is structured as follows:

1. In chapter 2, we describe related papers among 100 paper-reviewed research reports and articles about lung sound detection, feature extraction of image data, classification using image data and deep learning models, and machine learning models.

2. In chapter 3, we describe the dataset, and data type, and implemented deep and shallow learning model architecture.

3. In chapter 4, we describe experimented results and compare them with existing works using deep learning models to detect abnormalities in respiratory sound and the other result we have gotten from neutrino data and how it has generalized our proposed work.

4. Finally, we conclude our thesis work in chapter 5.

Chapter 2
Respiratory Sound classification using deep and machine learning: a systematic review

Summary: In this chapter, we reviewed research reports and articles. We describe the processes that have been attempted to detect and classify respiratory sound using shallow learning, Deep learning model.

Keywords: Deep and shallow learning models.

Organization: The rest of the chapter is structured as follows: In section 2.1 we describe the research articles that inspired us toward our proposed work and section 2.2 has all the works in a form of a table.

In section 2.3 we describe what's next.

2.1 Related work

Due to the global expansion of respiratory infections, it is imperative to make a quick diagnosis of the issue. The management of respiratory disease depends heavily on early detection and prevention. Though raw data can be used to detect and categorize anomalies, visualization offers more information, improving identification. Using AI, this identification process may be simplified. Doctors can help their patients more effectively by correctly identifying and classifying lung sounds using visualization with respiratory sounds. Lung sounds are nonstationary and non-linear signals, making them challenging to analyze and differentiate. With the use of an electronic stethoscope, automated analysis was made possible. The creation of algorithms that can recognize frequent aberrant breath sounds (such as wheezes and crackles) from clinical and nonclinical contexts was spurred in 2017 by the creation of the largest publicly accessible respiratory sound database. Generally speaking, respiratory sounds can be classed as normal or adventitious. Adventitious sounds are RS overlaid on typical respiratory noises, which can include crackles or wheezes. When lung fibrosis (fine crackles) or chronic airway obstruction are present in cardiorespiratory illnesses, crackles, which are discontinuous, explosive, and non-musical sounds that typically last less than 20 ms, frequently occur (coarse crackles). High-pitch ends with a duration of more than 100 ms are wheezes. Patients with obstructive airway disorders frequently have them can signify obstructive airway diseases, such as asthma and COPD, too. The dataset includes breathing cycles that were noted as wheezes by professionals who collected the data, either crackles or no strange sounds.
We divided earlier research into the following categories: Aykanat et al. [10] provided a convolutional network as well as a support vector machine-based solution for lung sound classification. The two feature extraction techniques are spectrogram production utilizing the short- time Fourier transform and Mel frequency cepstral coefficient (MFCC) feature extraction). They employed SVM along with MFCC

characteristics, which is a widely utilized method for categorizing audio. The mel frequency cepstrum (MFC), used in sound processing, is a representation of a sound's short-term power spectrum based on a linear cosine transform of a log power spectrum on a non- linear mel scale of frequency. The coefficients that make up an MFC are called MFCCs. They come from a particular cepstral interpretation of the audio clip. MFCC features are also employed in [11], where clips are first preprocessed by windowing and framing, and then MFCC features are extracted. Also, the second-level MFCC-2 feature values are computed to address the uneven and huge dimensionality issues in the following paragraphs. A spectrogram is a graphic depiction of the frequency spectrum in a sound or other signal as it varies over time or in response to other factors. They are widely utilized in the domains of seismology, speech processing, sonar, and radar. Since MFCC features are frequently employed in audio detection systems, the experiments that were conducted using these features allowed for the establishment of a baseline value for each of the following characteristics: accuracy, precision, recall, sensitivity, and specificity. Audio detection also uses spectrogram images. They were never put to the test using CNNs for respiratory audio, though. SciPy was used to build the MFCC datasets. They processed these datasets utilizing support vector machines. The open-source Pylab graph creation library and numerous open-source image processing packages were combined to create the spectrogram dataset. The experimenters altered the method to produce 28 28 grayscale spectrograms because the previous ones were 800 600 RGBA and would not fit in the computer's memory. This would allow CNN to process them. They tested four different situations using both of the suggested methods using a dataset of 17930 sounds from 1630 people. They reported an accuracy of 86 percent for classifying healthy-pathological conditions using both SVM and CNN. They came to the conclusion that the CNN and SVM machine learning algorithms can accurately categorize and pre-diagnose respiratory audio given the huge amount of data, and that spectrogram image classification with the CNN and SVM algorithms works as well as the SVM algorithm.

Acharya et al. [12] demonstrated a deep learning-based method for classifying lung sounds. Due to its unmatched performance in a range of applications, including clinical diagnostics and biomedical engineering, deep learning has received a lot of attention lately. These deep learning paradigms have the important benefit that the network learns usable features and abstract representations from the data through training, eliminating the need to manually create features from the input. They employed multiple data augmentation approaches to expand the dataset because it is very modest for training a deep learning model. These data augmentation techniques not only expand the dataset but also assist the network in learning usable data representations despite varying recording settings, equipment, patient age and gender, interpatient variability in breathing rate, etc. They employed a mel-frequency spectrogram with a window size of 60 ms and 50 percent overlap for feature extraction. Then, each breathing cycle is transformed into a 2D image, where each value represents the log amplitude value of the signal corresponding to that frequency and time window, and rows correspond to frequencies in mel scale and columns to time (window).They suggested a hybrid CNN-RNN model that has three stages: the first stage is a deep CNN model that extracts abstract feature representations from the input data, the second stage is a bidirectional long, short-term memory layer (Bi-LSTM), and the third stage has fully connected softmax layers that convert the output of the previous layers to class prediction. Due to the irregular character of wheeze and crackle as well as their temporal and frequency variance, these hybrid CNN-RNN designs have been more frequently used in sound event detection; nonetheless, comparable hybrid architectures may be effective for lung sound classification.

Deep learning models encountered a challenge because they needed a significant amount of data for training. They presented a patient-specific model-tuning strategy that can make use of deep learning

techniques even with a limited amount of patient data in order to solve these drawbacks of the existing approaches. In the suggested strategy, the deep network is initially trained on a sizable database to discover feature representations specific to a certain domain. The network is then retrained using the scant amount of patient-specific data that is available. As a result, they were able to convert the deep network's acquired domain expertise into patientspecific models, producing consistent and highly accurate predictions for the patient-specific class. They used training samples to train the three-stage network in their proposed model. The learned CNN-RNN stage weights are then locked in their pre-trained values for a new patient, while only the final stage is retrained with patient-specific breathing cycles. According to their research, the hybrid CNN-RNN model produced a score of 66.31 percent on an 80-20 split for the classification of the four classes of the respiratory cycle. To identify unhealthy patients, they then proposed a patient screening and model- tuning strategy. They then developed patientspecific re-training models, which significantly improved the reliability of the results for the original train-test split, achieving a score of 71.81 percent for leave-one-out cross-validation on the ICBHI17 dataset. The authors suggested feature extraction of lung sounds using wavelet coefficients and their classification by neural network and support vector machines in their publication, [13]

"Classification of Normal and Abnormal Lung Sounds Using Neural Network and Support Vector Machines." The study employed a total of 48 samples for the training and test. Lung sounds were divided into six groups. SVMs are an extremely good classifier for categorizing lung sounds, with an accuracy range of 93.51 to 100. On a dataset of 38 recordings, Pramono et al. [14] examined many characteristics to categorize wheezes and typical respiratory sounds. The dataset used in this study, which tested the discriminatory power of several feature types used in analogous studies in the past, included 38 recordings from various sources. There were 425 incidents total, 223 of which were wheezes, with the remaining events being typical. They showed that specific individual features, such as the tonality index and the MFCC, are far more effective at detecting wheezes. However, compared to more straightforward time-domain properties, their calculation requirements are greater. Furthermore, it has been demonstrated that after a certain number of characteristics, performance improvement is quite limited, even though using several features might sometimes increase classification accuracy. In their conclusion, they noted that while the classifier utilized in this study was somewhat basic, using other, more sophisticated classifiers, like support vector machines and artificial neural networks, could improve classification performance at the expense of increased computing complexity. As a result, it's critical to consider all conflicting needs when choosing a feature for wheezing detection across various applications. The outcomes of their experiments with various attributes are shown in [8].

Rao et al. [15] used acoustic methods to analyze the lungs. They discussed the acoustic features of various lung conditions. The physical makeup of the human thorax and methods for detecting breathing noises are also covered.

Along with several classifiers, the authors have also covered in detail the various signalprocessing methods needed to examine these noises. Bahoura and Pelletier [16] used cepstral features to distinguish normal and wheezing sounds. They worked with 12 instances from each class and reported the highest true positive value of 76.6 percent for wheezing sounds. They also reported 90.6 percent true positives for normal sounds with Fourier transform- based features. Demir et al. [17] employed a CNN-based method to classify lung

sounds from the ICBHI 2017 dataset. For the extraction of deep features, they presented new pre-trained Convolutional Neural Network (CNN) models like VGG16 and

AlexNet. However, as these CNN models haven't been trained on sound datasets, sound features aren't adequately captured. As a result, lung sound-based spectrogram pictures were used to train the suggested CNN model. Additionally, the parallel-pooling structure was used in the suggested CNN design to improve classification performance. To improve classification performance, an averagepooling layer and a max-pooling layer are coupled in parallel in the CNN design. The Linear Discriminant Analysis (LDA) classifier is fed data from the deep features using the Random Subspace Ensembles (RSE) technique. They reported an overall accuracy of 71.1 percent and a maximum accuracy of 83.2 percent for the healthy class.

Ma et al. [12] developed a method to differentiate respiratory sounds using the non-local block in the ResNet architecture. In order to attain the best state-of-the-art accuracy, they suggested a LungRBN model, which combines wavelet feature extraction and short-time Fourier transform (STFT) with a product of two ResNet models through a fully linked layer. Discovering strategies in order to automatically supplement current data, however, has received less focus than finding techniques to improve detection accuracy significantly. They introduced LungRN+NL, an enhanced adventitious lung sound classification, which combines a mix-up data augmentation method with a non-local layer of the ResNet neural network to address this issue. In order to extract features from lung sounds, we opt for the time-frequency analysis technique known as short-time Fourier transform (STFT), taking into account the crucial differentiation between various categories. On the basis of experiments, the ICBHI 2017 dataset has an accuracy of 52.26 percent.

In "Investigating into segmentation methods for diagnosis of respiratory diseases using adventitious respiratory sounds," [46] the author wanted to investigate the segmentation methods to diagnose respiratory diseases using respiratory sounds. They used IMF feature extraction techniques and Random forest, EMD classifier for identifying respiratory conditions. They have used the ICBHI17 dataset for classification. The results show the accuracy as 0.88, the F1 score as 0.81, and specificity as 0.91. Dokur [46] employed a rectangular window made from a single respiratory sound (RS) cycle, which was windowed time samples that were later normalized. The adjusted RS signal is split into 64 samples of lengthy segments before the features are extracted. The power spectrum of each segment is then calculated, and its components are summarized in sync. The averaged power spectrum components result in 32dimensional feature vectors, which are generated. For the classification of nine different RS classes—bronchial sounds, Broncho vesicular sounds, vesicular sounds, crackles sounds, wheezes sounds, stridor sounds, grunting sounds, squawk sounds, and sounds of friction rub— this study compares the classification performances of three different types of networks: multilayer perceptron (MLP), grow and learn (GAL), and a novel incremental supervised neural network (ISNN). They conducted a three-stage study of respiratory sounds, including feature extraction, normalization, and categorization of the respiratory sounds using artificial neural networks (ANNs). One cycle of RS is confined within a rectangular window that is created in the first stage. The window has 8,192 samples in it. The windowed time samples are then normalized, setting the respiratory signal power in the window to 1. Using the window's normalized data, feature vectors are created in the second stage. The classification of the RSS is carried out using artificial neural networks in the final stage, and in this study, multi-layer perceptrons were used. This study reported an accuracy of 92%.

In “A Deep Convolutional Neural Network in a Wearable Cough Detection System.”[47] The author wanted to design a wearable cough detection system using a convolutional neural network. They have created a database of 627 sounds from healthy and non-healthy people. With MFCC feature extractor and CNN and achieved a sensitivity of 95.1% and a specificity of 99.5%.

Using a CNN-based classification method, Shivakumar [48] divided respiratory sounds into two categories: crackles and wheezes. After pre-processing the audio samples, they built a neural network using modified CNNs to build the dataset's base model. Later, they employed an Adam optimizer with a 64-batch batch size and a 0.009 learning rate. For the first model, the author ran the model on wheezes and crackles individually for 10 epochs after using both wheezes and crackles simultaneously for 10 epochs. The outcomes for both a 90-10 and an 80-20 train-test split were identical. The author also showed how it is quite advantageous to divide the sounds into various models. In this study, two models were put forth, and they generated test accuracies of 50% and 100%, respectively.

On the ICBHI 2017 dataset, Faustino [49] reported a CNN-based method for the detection of wheeze and crackle. MFCC and power spectral density values were extracted from the audio clips for the investigation. CNN received these for classification. They discovered that using a Mel Spectrogram with a Convolutional Neural Network architecture to classify lung sounds is more advantageous than doing so using MFCC characteristics. These results, however, did not outperform those from the other study, which also used the same dataset but employed an RNN architecture with MFCC features. These results suggest that, for the categorization of lung sounds, using a Recurrent Neural Network architecture in conjunction with MFCCs is preferable to using a convolutional-based technique. The MFCC approach works better when paired with an RNN than a CNN because it uses the discrete cosine transform to compress and correlate the signal characteristics. A CNN architecture makes inefficient use of the MFCCs since it capitalizes on local trends in the data. With the temporal context of the data and access to all input features without the need for shared parameters, an RNN is constructed using an FNN as the interior network, making it a far better architecture for decoding MFCC input. Finally, 43% test accuracy was reported utilizing a fivefold cross-validation method.

To discriminate between healthy and non-healthy cases, Kok et al. [50] used a number of indicators, including MFCC, DWT, and temporal domain measures. The Wilcoxon Rank Sum statistical test was applied to ascertain the significance of the retrieved characteristics after a number of features were investigated. The feature combination that gave the least amount of redundancy and the most amount of relevance was then found using a feature selection algorithm based on mutual information using the significant features as input. The cases were categorized as random using the sampling and boosting methods. Their accuracy, specificity, and sensitivity scores were 87.1%, 93.6%, and 86.8% respectively.

In “Efficient FPGA-based architecture of an automatic wheeze detector using a combination of MFCC and SVM algorithms” [51], the author wanted to detect the wheeze sounds using the MFCC features and SVM classifier. They have used two databases each of 12 sounds. One of them has healthy sounds and the other has asthmatic sounds. The accuracies using XSG, optimized XSG and Matlab are 0.9372, 0.9359, and 0.9359.

The author wanted to classify the lung sounds using a Convolutional Neural Network (CNN) and 12 MFCC coefficients in “Lung sounds classification using convolutional neural networks.” [52]. They have used the

RALE data set of lung sounds and 50 more recordings from the respiratory acoustics laboratory of the University of Manitoba in Winnipeg, Canada. The results of the classification show 93.26% of accuracy.

2.2 State-of-the-art works and their performances

Table 1. State-of-the-art works and their performances.

Author	Dataset	Performance (ACC, AUC, SEN, SPEC)
Aykanat et al. [10]	Electronically recorded (17,930)	86%,–, 86%, 86%
Acharya et al. [12]	ICBHI'17 dataset(920)	96%, , 48.63%, 84.14%
Pramono et al. [14]	Multiple repositories (38)	, 89.19%, 83.86%, 81.19%
Mukherjee et al. [11]	Electronically recorded (17,930)	86%, , ,
Rao et al. [15]	Multiple sources	93% – 95%, , ,
Demir et al. [17]	ICBHI'17 dataset(920)	71.15%, , ,
Ma et al. [12]	ICBHI'17 dataset(920)	, , 41.32%, 63.2%
L. et al. [45]	ICBHI'17 dataset(920)	88%, , , 91%
Dokur. [46]	Individual patient data and RALE (180)	98%, , ,
Amoh et al.[47]	database of 627 sounds	, , 95.1%, 99.5%
Shivakumar [48]	ICBHI'17 dataset(920)	100%, , ,
Faustino[49]	ICBHI'17 dataset(920)	43%, 51%, 36%
Kok et al.[50]	ICBHI'17 dataset(920)	87.1%, , 86.8%, 93.6%
Boujelben et al. [51]	Dataset of 12 sounds	93.72%, , 93.72%, 93.72%
Bardou et al. [52]	RALE dataset+ 50 more recordings	93.26%, , ,

2.3 What's next?

The rest of our thesis work is structured as follows:

1. In chapter 3, we describe the respiratory dataset, the data type's description, all model architecture, and the implementation
2. In chapter 4, we have described the result and analysis and compared them with state-ofart works
3. Finally, we conclude our thesis work in chapter 5.

Chapter 3
Respiratory dataset, data types description, all model architecture, and implementation

Summary: In this chapter, we give a details overview of dataset collection, data type creation, all the model architecture, and the implementation details of all the model architecture.

Key topics: dataset details and data types, model architecture, and implementation.

Organization: The rest of the chapter is structured as follows: In section 3.1 we describe the dataset, Section 3.2 includes data type description, and in section 3.3 all the deep learning and machine learning model architecture, and implementation. In section 3.4 we describe what's next

3 Dataset, Data type, and model architecture

This section includes dataset collection, data type creation, model architecture, and implementation.

3.1 Dataset

To develop of a robust system, it is important to ensure that the dataset mimics real-world problems. Our system was trained on a publicly available respiratory sound database [19] [20], which is associated with the International Conference on Biomedical and Health Informatics (ICBHI). Most of the database consists of audio samples recorded by the School of Health Sciences, University of Aveiro (ESSUA) research team at the Respiratory Research and Rehabilitation Laboratory (Lab3R), ESSUA, and at Hospital Infante D. Pedro, Aveiro, Portugal.The second research team, from the Aristotle University of Thessaloniki (AUTH) and the University of Coimbra (UC), acquired respiratory sounds at the Papanikolaou General Hospital, Thessaloniki and at the General Hospital of Imathia (Health Unit of Naousa), Greece. The Respiratory Sound Database contains audio samples, collected independently by two research teams in two different countries, over several years. The database consists of a total of 5.5 hours of recordings containing 6898 respiratory cycles, of which 1864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes, in 920 annotated audio samples from 126 subjects [21].

To collect data, disparate stethoscopes and microphones were used. The audios were recorded from the trachea and 6 other chest locations: left and right posterior, anterior, and lateral. The audio was collected in clinical and non-clinical settings from adult participants of different ages. Participants encompassed patients with lower and upper respiratory tract infections, pneumonia, bronchiolitis, COPD, asthma, bronchiectasis, and cystic fibrosis. The ICBHI database consists of 920 audio samples from 126 subjects. These are annotated by respiratory experts and used as a benchmark in the field. Each respiratory cycle in the dataset is annotated among 4 classes. The annotations basically cover 2 broad groups: healthy and non-healthy. The nonhealthy category is further divided into wheeze and crackle with some cycles having both issues. Among 6898 cycles totaling to 5.5 hours, 1864 cycles have crackles while 886 have wheezes. There are 506 cycles, which have both wheezes and crackles. While recording, the participants were seated. The acquisition of respiratory sounds was performed on adult and elderly patients. Many patients had COPD with comorbidities (e.g., heart failure, diabetes, and hypertension).

Further, noise exists, such as the rubbing sound of the stethoscope with the patient's dress, and background talking. Such varieties in the data made it challenging to identify problems in the respiratory sounds. One of the most challenging aspects of the audio clips was the presence of a heartbeat sound along with the breath sounds. No preprocessing was performed to remove the heartbeat sounds.

Table 2. Respiratory sound database [19]

Clip type	Number of clips
Healthy	3642
Non-healthy	3256

While recording, the participants were seated. The acquisition of respiratory sounds was performed on adult and elderly patients. Many patients had COPD with comorbidities (e.g., heart failure, diabetes, and hypertension). Further, noise exists, such as the rubbing sound of the stethoscope with the patient's dress, and background talking. Such varieties in the data made it challenging to identify problems in the respiratory sounds. One of the most challenging aspects of the audio clips was the presence of a heartbeat sound along with the breath sounds. No preprocessing was performed to remove the heartbeat sounds.

3.2 Data Type

In our experiment, a total of six different categories of data were used. The data types include raw data, MFCC, spectrograms, zero crossing rates, spectral centroid, and spectral roll-off. Here is a basic introduction to them.

Raw Data: In order to analyze and discover relationships between many entities, feature extraction is a key step. The models cannot directly interpret the audio data presented, so feature extraction is utilized to transform it into a format that can be understood. It is a method that explains most of the material in a clear manner. Algorithms for classification, prediction, and recommendation all require feature extraction [29].

The audio signal is a three-dimensional signal in which three axes represent time, amplitude, and frequency.

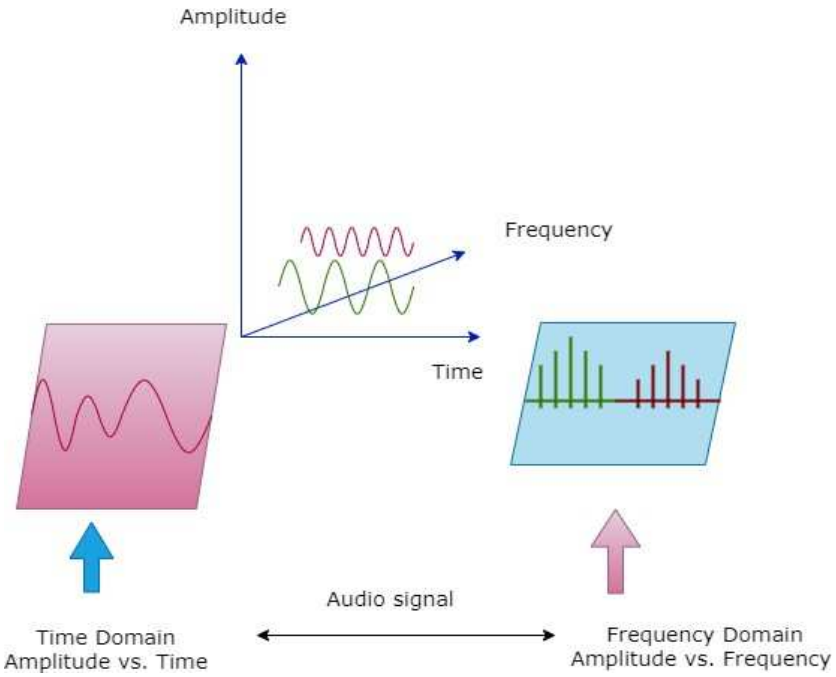

Fig. 2. Audio signal is a three-dimensional signal in which three axes represent time, amplitude, and frequency.

Wave plots show us how loud the audio is at any particular moment. A spectrum displays the many frequencies present at any given time together with their amplitude. The properties of sound that are crucial and particular to each audio file are amplitude and frequency. In figure 1 an amplitude vs. time waveform is plotted using a wave plot, where the first axis is amplitude and the second is time.

Data is transformed into a short-term Fourier transform using.stft. STFT transforms signals so that we can determine the frequency's amplitude at a particular moment. We may use STFT to calculate the amplitude of different frequencies that are playing at a specific time in an audio source. The figure 2 type of figure from our dataset was generated using the same procedure.

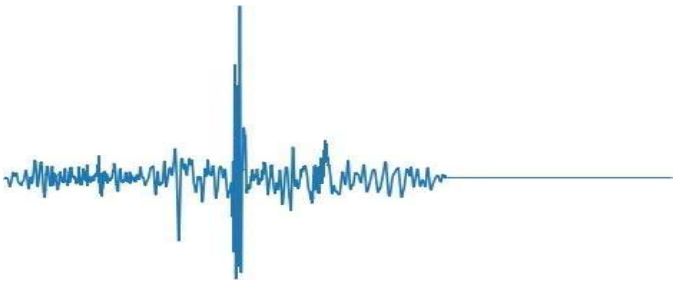

Fig. 3. Raw data representation of a dataset

MFCC: A representation of a sound's short-term power spectrum used in sound processing is called a Mel-frequency cepstrum (MFC), which is based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. An MFC is made up of a number of coefficients known as mel-frequency cepstral coefficients (MFCCs) [30]. They are derived from an audio clip's cepstral representation (a nonlinear "spectrum-of-a-spectrum"). The Melfrequency cepstrum (MFC) differs from the cepstrum in that the frequency bands are evenly spaced on the mel scale, which more closely resembles the response of the human auditory system than the linearly-spaced frequency bands used in the conventional spectrum.

Fig. 4. MFCC sample from our dataset

When used in audio compression, for instance, this frequency warping can improve the representation of sound and potentially lower the transmission bandwidth and storage needs of audio signals. When working with audio signals, this feature is one of the most crucial ways to extract a feature from the audio signal. A signal's Mel frequency cepstral coefficients (MFCCs) are a small group of characteristics (often 10–20) that succinctly define the general contours of a spectral envelope. We can determine how many mfccs are calculated on how many frames by printing the shape of the mfccs. The first value is the total number of mfccs calculated, and the second value is the total number of available frames.

Spectrogram: A spectrogram is a visual representation of the "loudness" or signals strength over time at different frequencies contained in a specific waveform. One may observe the amount of energy at different frequencies, such as 2 Hz vs. 10 Hz, as well as how it changes with time [31]. It is a visual representation of the sound or other signal's spectrum of frequencies as they change over time.

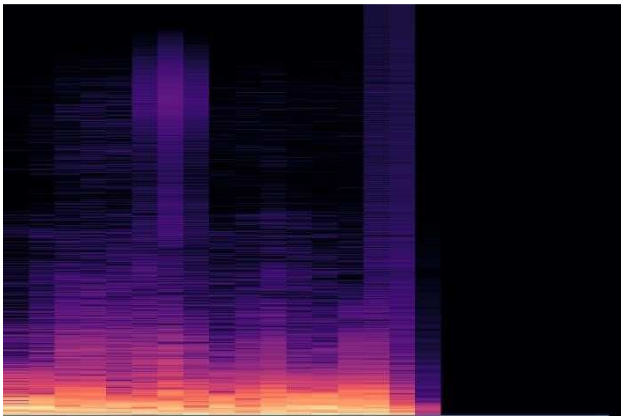

Fig. 5. Spectrogram sample of the dataset

It shows how the frequencies for specific musical signals change over time. Data is transformed into a short-term Fourier transform (stft). STFT transforms signals so that we can determine the frequency's amplitude at a particular moment. .specshow is used to display the spectrogram, which may be used to calculate the amplitude of different frequencies playing at a specific moment in an audio source using STFT. Figure 4 represents the spectrogram sample of the dataset.

Zero crossing rate: One fundamental aspect of an audio signal that is frequently used in audio classification is zero crossings. Zero crossings provide a rough calculation of the spectral centroid (SC) and dominant frequency [35]. The zero-crossing rate (ZCR) is the rate at which a signal changes from positive to zero to negative or from negative to zero to positive [32].

The ZCR, which is defined as the number of zero crossings in the temporal domain within one second, is one of the most affordable and straightforward features. The dominant frequency of a signal is measured by the ZCR, claims Kedem [38]. Due to its simplicity, ZCR is a well-liked feature for speech/music differentiation [36, 37]. It is, nevertheless, widely applied in a variety of other audio application fields, including musical genre classification [44], highlight detection [42], speech analysis [44], singing voice detection in music [40], and environmental sound recognition [41]. Its value has been widely used in both speech recognition and music information retrieval, being a key feature to classify percussive sounds [33]. Another way to define it is the rate at which a signal's signs change is known as the zero crossing rate. Both speech recognition and the retrieval of music information have made extensive use of this characteristic. For extremely percussion-heavy sounds, such as those found in metal and rock, it typically has higher values.

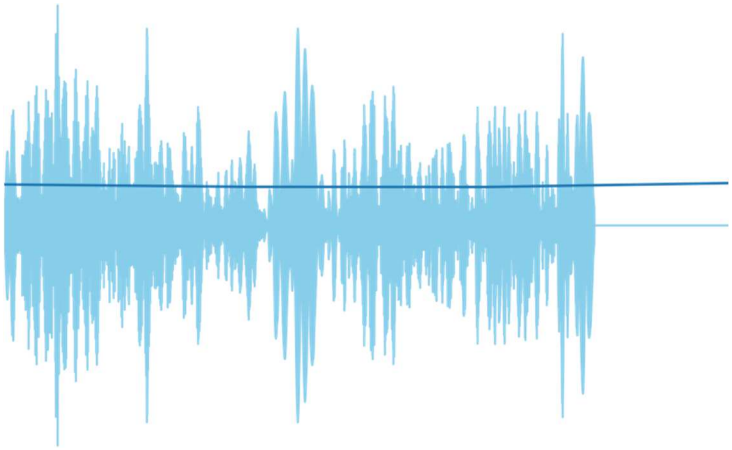

Fig. 6. Sample of zero crossing rate of the dataset

It is determined as the weighted mean of the frequencies present in the sound and identifies the location of the "center of mass" for a sound. If the audio's frequencies remain constant throughout, the spectral centroid would be in the middle, and if the sound has high frequencies at the end, it would be near the end. Similar to the zero crossing rate, the spectral centroid appears to falsely increase at the start of the signal. The reason for this is that because the initial quiet has such a modest amplitude, high-frequency components have a chance to predominate.

Zero crossing rate is computed as:

a. Let, (n), n = 0,1,…N–1 be the samples of the ith frame.

b. Zero crossing rate of each frame is calculated as:

$$Z(i) = \sum_{n=0}^{N-1} |sgn\,[x_i(n)] - sgn[x_i(n-1)]|$$

Where,

$$sgn\,[x\,(n)] = \begin{cases} 1, x^i(n) \geq 0 \\ -1, x_i\,(n) < 0 \end{cases}$$

Spectral Centroid: The spectral centroid is a simple measure of spectral position and shape. The spectral centroid is the center of the 'gravity' of the spectrum [34]. It is determined as the weighted mean of the frequencies present in the sound and identifies the location of the "center of mass" for a sound. If the audio's

frequencies remain constant throughout, the spectral centroid would be in the middle, and if the sound has high frequencies at the end, it would be near the end. Similar to the zero crossing rate, the spectral centroid appears to falsely increase at the start of the signal. The reason for this is that because the initial quiet has such a modest amplitude, high-frequency components have a chance to predominate. It is computed as follows:

a. Let (n), n = 0, 1,…N–1 be the sample of the ith frame, with x_i(k), k = 0,1,...N–1 as the discrete Fourier transform (DFT) coefficients of the sequence.

b. Compute the SC of each frame as:

$$C(i) = \frac{\sum_{k=0}^{N-1} k|x_i(k)|}{\sum_{k=0}^{N-1} |x_i(k)|}$$

For each epoch of the frame of the audio signal, the SC feature can be the average value of the SC over all frames.

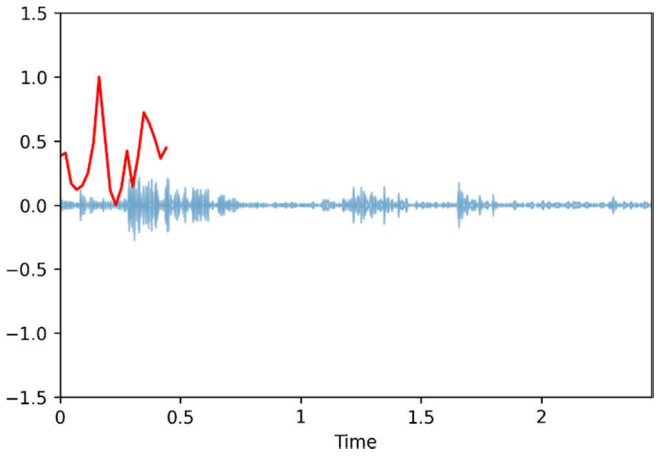

Fig. 7. Sample of spectral centroid of the dataset

Spectral roll-off: The frequency below which a certain amount of the total spectral energy, for example, 85%, lies is known as the spectral roll-off. Results are also provided for each frame. The frequency below which a specific portion (often 80%–90%) of the magnitude distribution of the spectrum is concentrated in the spectrum is known as the spectral roll-off.

Fig. 7. Sample of spectral roll-off of the dataset

The computation looks like this:

a. Assume that the discrete Fourier transform (DFT) coefficients of the sequence are represented by Xi(k), k = 0,1,...N-1, and that xi(n), n = 0,1,...N-1 is the sample of the ith frame.

b. Calculate the spectral roll-off for the sample that fulfills.

In some cases, the P parameter is set between 80 and 100.

$$\sum_{k=0}^{k(i)} |x_i(k)| = \frac{P}{100} \sum_{k=0}^{N-1} |x_i(k)|$$

3.3 Model Architecture and Implementation

Both machine learning and deep learning models have been applied. Machine learning models include logistic regression and support vector machines. As deep learning models, ResNet50, VGG16, InceptionV3, and MobileNet are employed. Here are a few summaries of them.

ReseNet50: He et al. [20] proposed the ResNet 50, a Residual Network with 50 layers. This model's input size is fixed at 224×224, and its convolutional layers are the same size as those in VGG networks, which follow some straightforward designs like. The outcome is the same for layers with the same amount of filters. If the convolved output size is cut in half while maintaining the time complexity layer, the number

of filters is increased by a factor of two. The model's final layers are a 1000-way fully connected layer with softmax and an average pooling layer. In comparison to VGG nets, this model has fewer filters and is less complex, albeit there are other variations like ResNet101 and ResNet152. The ResNet50 configuration layers are shown in Fig. 2. A network can accept an image as input if its height, width, and channel width are all multiples of 32.

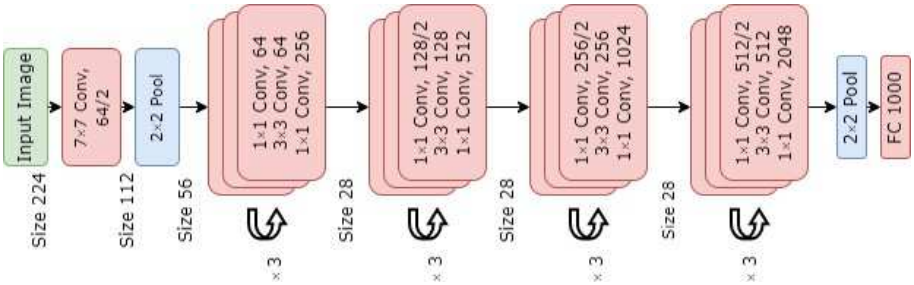

Fig. 6. ResNet50 architecture

The size of the input is 224×224×3. Using 77 and 33 kernel sizes, respectively, each ResNet architecture conducts the first convolution and max-pooling. The network's Stage 1 then begins, and it consists of 3 Residual blocks with a total of 6 layers. All 3 layers of the block in stage 1's first stage were constructed using kernels that were 64, 64, and 128 in size, respectively.

The identity relationship is represented by curved arrows. The size of the input will be cut in half in terms of height and width but doubled in terms of channel width since the convolution operation in the residual block is carried out with stride 2. The channel width doubles and the input size is cut in half as it moves from one step to the next. Bottleneck design is applied to deeper networks like ResNet50, ResNet152, etc. A total of three layers are piled on top of one another for each residual function F. Convolutions (1, 3, and 1) make up the three levels. The reduction and subsequent restoration of the dimensions are accomplished by the 11 convolution layers. With lower input and output dimensions, the 33 layers is left as a bottleneck. The network's final layer is an average pooling layer, which is followed by a layer of 1000 neurons that is fully connected (ImageNet class output).

VGG16: VGG Architecture [27]: A dimensioned image Fig. 7 serves as the network's input (224, 224, 3). The first two layers include the same padding and 64 channels with a 3×3 filter size. Following a max pool layer of stride (2, 2), two layers have convolution layers of 128 filter size and filter size (3, 3). A max-pooling stride (2, 2) layer that is the same as the layer preceding it comes next. Then, 256 filters with filter widths of 3 and 3 are distributed over 2 convolution layers. After that, there are two sets of three convolution layers, and then a max pool layer comes next. Each filter contains 512 filters and the same padding with filter size (3, 3).

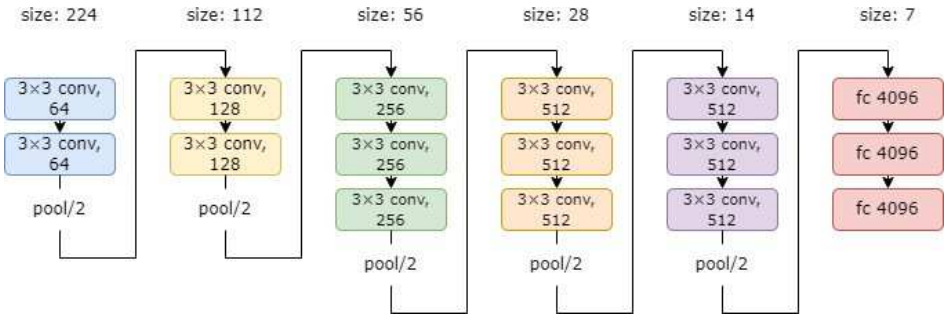

Fig. 7. VGG16 architecture

This image is then applied to the stack of two convolution layers. These convolution and max- pooling layers both use 3×3-sized filters. In order to change the number of input channels, it additionally uses 1×1 pixels in some of the layers. After each convolution layer, 1 pixel (the same padding) is inserted to prevent the image's spatial characteristic. After adding a convolution and max-pooling layer to the stack, we got a (7, 7, 512) feature map. In order to construct a (1, 16384) feature vector, this output is flattened. The following three layers are all fully interconnected; the first layer outputs a (1, 256) vector using the most recent feature vector as input, the second layer also generates a (1, 128) vector, and the third layer generates 1000 channels for two classes. Every hidden layer uses ReLU [22] as its activation function. ReLU is more computationally effective since it results in speedier learning and lowers the possibility of vanishing gradient problems. The list of VGG topologies in Fig. 3 is extensive. VGG-16 has two unique versions, as can be shown (C and D).

We obtained a (7, 7, 512) feature map after adding a convolution and max-pooling layer to the stack. This output is flattened to create a (1, 16384) feature vector. The next three layers are all completely connected; the first layer uses the most recent feature vector as input and outputs a (1, 256) vector, the second layer also produces a (1, 128) vector, and the third layer produces 1000 channels for two classes. The only difference between them is the use of the (3, 3) filter size convolution in place of some convolution layers (1, 1). These two have, respectively, 134 million and 138 million attributes.

MobileNet: Initially, MobileNet uses a simple 2D convolution layer. Then, a set of convolution layers with varying strides and filter counts known as Depth- wise Separable are attached one after the other. In Fig. 4 following the input, each convolutional block proceeds in the following order: BatchNormalization, ReLU activation, and then it is passed to the following block. 32 filters with a stride of 2 and a kernel size of 3x3 make up the first convolution block. As previously stated, a BatchNormalization layer and a ReLU activation come next. Then comes the Depth-wise Separable convolution layer, which is the foundational building component of the MobileNet design. Depth- wise convolution and then point wise convolution make up this procedure. Unlike depth-wise convolution, which applies a kernel to each channel, normal convolution applies a block of the same channel as the picture to all of the channels.

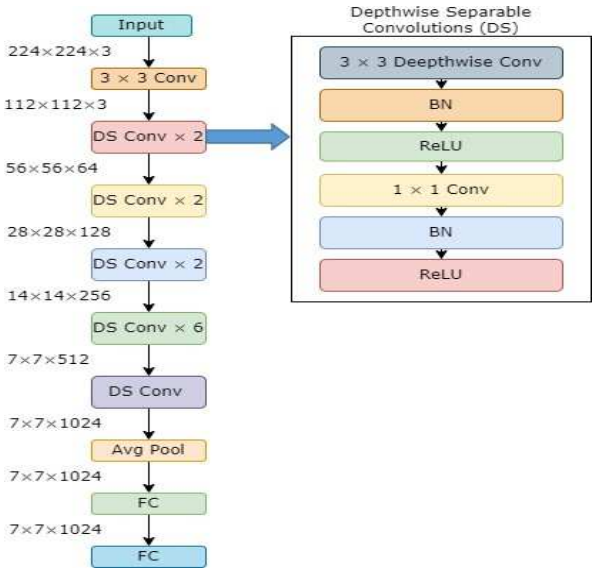

Fig. 8. MobileNet architecture

Then, a point wise convolution 1x1 conv is applied to the stacked output layers. The depth-wise convolution and point-wise convolution functions are referred to collectively through a combination layer function. The point-wise convolution is applied n times and is computationally less expensive than performing n transformations on images.

For an n-filter number, depth-wise convolution with a stride of 2 reduces the size first, then stride 1 is used for depth-wise convolution. With each combination layer, the number of channels gradually increases from 32 to 1024. Iteratively, the block is called five times at channel output 51Global Average Pooling comes last, followed by the final output layer. The classes are to be listed in the output layer, which is a dense layer. It must be dense if the classes are 3. Softmax is the activation method employed.

Inception-V3 architecture: Szegedy et al. [22] [23] introduced the Inception-v3 convolutional neural network type as GoogLeNet. This network has 48 layers and can classify photos into 1000 different categories. This model's fixed input size is 299×299 pixels. This model is built on a multiscale method that combines various classifier structures with different backpropagation sources. The Inception-v3 model expands the network's span and depth without imposing costs. To enable generating more complex decisions, this model applies many Inception layers in convolution on the input feature map at various sizes. The architecture of Inception-v3 is shown in Fig 5. Convolutions like 5×5, which significantly reduce the input dimensions, were occasionally used in Inception V1. The neural network's accuracy suffers as a result of this. This is because if the input dimension is reduced too much, the neural network is vulnerable to information loss. Additionally, compared to 33, the complexity decreases when larger convolutions like 55 are used. An asymmetric 13 convolution followed by a 31 convolution can be used to factorize a 33 convolution. This is the same as sliding a two-layer network with a 33 convolution receptive field, but it is 33% less expensive.

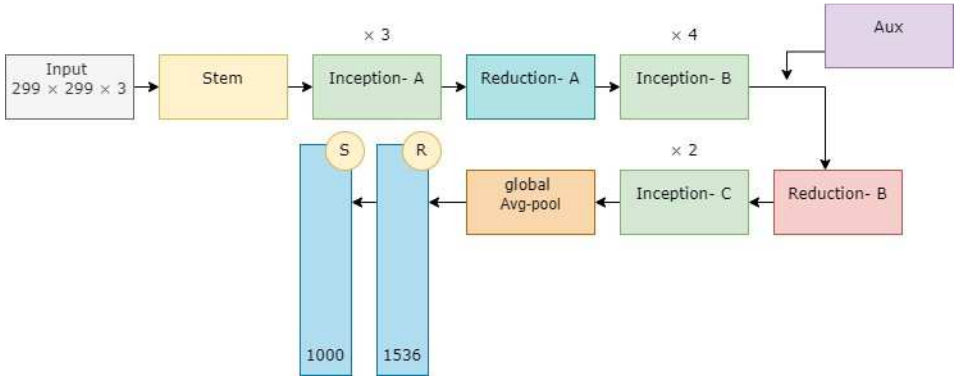

Fig. 9. Inceptionnet-V3 architecture

Only when the input size is m×m does this factorization function well for early layers when input dimensions are large (m is between 12 and 20). The auxiliary classifier enhances the network's convergence, per the Inception V1 architecture. in the architecture of Inception V2. The two 33 convolutions take the place of the 55 convolutions. Due to the fact that a 55 convolution is 2.78 more expensive than a 33 convolution, this also reduces calculation time and hence boosts computation speed. Therefore, using two 3×3 layers rather than five 5×5 layers improves architecture performance. The use of RMSprop optimizer, batch normalization in the fully connected layer of the auxiliary classifier, and Use of 7×7 factorized are all additional characteristics of Inception V3, which is identical to and includes all the features of Inception V2. Another technique for regularizing the classifier is label smoothing regularization, which calculates the impact of label dropout during training. It stops the classifier from making excessively confident predictions about a class. The addition of label smoothing improves the error rate by 0.2%.

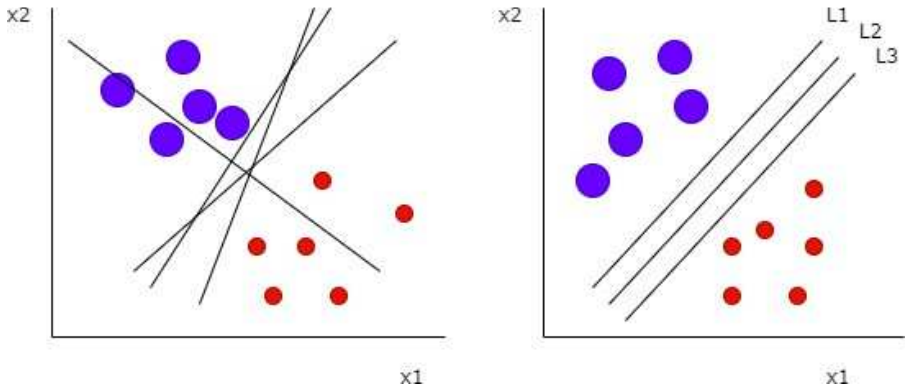

Fig. 10. Linearly Separable Data points

It is extremely obvious from the above graphic that there are numerous lines (our hyperplane in this case is a line as we are only taking into account two input features, x1, x2) that separate our data points or perform a classification between red and blue circles. The line that shows the greatest separation or margin between the two classes is the best line, or generally the best hyperplane, that divides our data points.

What's next?

The next chapter contains the result and analysis of lung sound visualization using several deep learning models.

Chapter 4
Results and Analysis

Summary: In this chapter, we compare the deep learning and machine learning results using different data types. We are also comparing our proposed work with existing work.

Key topics: result and analysis

Organization: The rest of the chapter is structured as follows: In section 4.1 we represent the obtained result, evaluation matrices, and analysis of the results with existing work and in section 4.2 we describe what's next

4.1 Result and Evaluation:

To measure the performance, we computed Accuracy (ACC), Precision (PREC) Sensitivity (SEN), and Area under the Curve (AUC). ACC, Precision, and SEN are computed as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision(PREC) = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Where TP, FP, TN and, FN refer to true positive, false positive, true negative and, false negative, respectively. In Table 3, we provide comprehensive results in terms of the aforementioned evaluation metrics (ACC, PREC, AUC, and SEN). We have applied VGG16, ResNet50, InceptionV3, and MobileNet on each data type and determined the result. Spectrogram performed better than raw data for all four methods. ResNet50 for MFCC and spectral centroid has the second highest AUC of 0.77.

Table 3. Performance evaluation on Raw Data input for Shallow Learning Models

Model name	Acc	Precision	Recall	F1
SVM	0.558	0.558	0.548	0.553
Logistic Regression	0.544	0.539	0.536	0.529

Table 4. Performance evaluation on ICBHI 2017 dataset using four pre-trained deep learning models.

Types of Data	Model name	ACC PREC AUC SEN
Wave	VGG16	0.4697 0.4697 0.4828 0.4697
	ResNet50	0.5877 0.5882 0.5637 0.7207
	MobileNet	0.5450 0.5659 0.54554 0.6460
	InceptionV3	0.5355 0.5355 0.5000 1.0000
MFCC	VGG16	0.4545 0.4545 0.5315 0.4545
	ResNet50	0.6090 0.6900 0.7700 0.7500
	MobileNet	0.5308 0.5288 0.5150 0.9910
	InceptionV3	0.5261 0.5261 0.5000 1.0000
Spectrogram	**VGG16**	**0.7232 0.7486 0.7930 0.6248**
	Resnet50	0.6558 0.6236 0.7168 0.6876
	MobileNet	0.5592 0.5581 0.5421 0.8496
	InceptionV3	0.5355 0.5355 0.5000 1.000
Spectral centroid	VGG16	0.5655 0.5455 0.6233 0.5500
	Resnet50	0.5735 0.5533 0.7700 0.9820
	MobileNet	0.5592 0.5581 0.5421 0.8496
	InceptionV3	0.5355 0.5355 0.5000 1.0000
Spectral roll-off	VGG16	0.4091 0.4091 0.3836 0.4091
	Resnet50	0.4787 0.5031 0.4680 0.7387
	MobileNet	0.5355 0.5355 0.5000 0.1000
	InceptionV3	0.4882 0.5163 0.3673 0.6991

Zero-crossing rate	VGG16	0.4697 0.4697 0.4828 0.4697
	Resnet50	0.5261 0.5261 0.3973 1.000
	MobileNet	0.3886 0.3749 0.3749 0.1150
	InceptionV3	0.5355 0.5355 0.5000 1.0000

In Table 3, we observed that 2D data representations worked better as compared to 1D data representation, which we call raw respiratory sound (wave). Of all 2D data representations, Spectrogram performed better and VGG16 provided an accuracy of 0.7232, a precision of 0.7486, an AUC of 0.7930, and a sensitivity of 0.6248.

We compared our findings with previous works that used the exact same dataset. Since they have used different metrics, we aimed at looking at whether they can be compared. In their works, they used an additional score - an average score of specificity and sensitivity.

On the whole, the proposed work still performs better.

Table 5. Performance comparison

Authors	PREC	AUC	SEN	SPEC
Ma et al (2020) [18]	-	-	41.32%	63.20%
Chambres et al (2018) [24]	-	-	20.81%	78.05%
Kochetov et al (2018) [25]	-	-	58.43%	73.00%
Acharya and Basu (2020) [26]	-	-	48.63%	84.14%
Ma et al (2019) [27]	-	-	31.12%	69.20%
Proposed work	**74.86%**	**79.30%**	**62.48%**	-

What's next?

The next chapter contains the conclusion of our research.

Chapter 5
Conclusion

In this work, we have analyzed deep visual features from 2D data representation of the respiratory sound to detect evidence of lung abnormalities. The primary motivation behind this was to study how well 2D representation and/or visual cues work in decision-making as opposed to 1D raw data. In our work, we have analyzed deep visual features using our Convolutional Neural Networks (CNN) tailored Deep Learning Models, we have used VGG16,
ResNet50, MobileNet, and InceptionV3, where 2D data in different formats such as Spectrogram, Mel-frequency Cepstral Coefficients (MFCC), spectral centroid, and spectral rolloff were considered. Our experiments on the publicly available respiratory sound database named ICBHI 2017 (5.5 hours of recordings containing 6898 respiratory cycles from 126 subjects) have proved that 2D data representation could help better understand/analyze lung abnormalities as compared to 1D data. Using the pre-trained deep learning model (VGG16), we achieved an AUC of 0.79 from Spectrogram as opposed to 0.48 AUC from the raw data. We also used a comparable technique using MFCC neutrino data and raw neutrino data, and as a consequence, we were able to detect neutrinos far more effectively utilizing picture data than raw data. This work can be a first step for AI-guided apps to detect aberrant respiratory sounds utilizing images created from the sound once improved accuracy and additional analysis are achieved. It can aid in the early diagnosis of various diseases as part of the comprehensive study. In the future, it may be possible to detect lung disorders by combining several data kinds. Additionally, using the ensemble method might improve accuracy.

References

[1] Tea Lallukka, Anoushka Millear, Amanda Pain, Monica Cortinovis, and Giorgia Giussani. Gbd 2015 mortality and causes of death collaborators. global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the global burden of disease study 2015 (vol 388, pg 1459, 2016). Lancet, 389(10064):E1–E1, 2017.

[2] Global status report on noncommunicable diseases 2014. geneva, world health organization, 2014. Available from: http://www.who.int/nmh/publications/ncd-status-report2014/en/.

[3] Lung Transplant. Lung Transplant | Michigan Medicine. (n.d.). Retrieved November 28, 2022, from https://www.uofmhealth.org/conditions-treatments/transplant/adult-lung-transplant?fbclid=IwAR2qYvOWZcIqsGy6x8iHS-K8PN9KO-jXAiutQz4ez_9Z7DvFGPT1FqiZr50

[4] KC Santosh. Speech processing in healthcare: Can we integrate? In Intelligent Speech Signal Processing pages 1–4. Elsevier, 2019.

[5] Himadri Mukherjee, Subhankar Ghosh, Shibaprasad Sen, Obaidullah Sk Md, KC Santosh, Santanu Phadikar, and Kaushik Roy. Deep learning for spoken language identification: Can we visualize speech signal patterns? Neural Computing and Applications, 31(12):8483–8501, 2019.

[6] Himadri Mukherjee, Sk Md Obaidullah, KC Santosh, Santanu Phadikar, and Kaushik Roy. Line spectral frequency-based features and extreme learning machine for voice activity detection from the audio signal. International Journal of Speech Technology, 21(4):753–760, 2018.

[7] Fred Hohman. Visualization in deep learning, Mar 2019.

[8] Sanket Doshi. Extract features of music, Apr 2019.

[9] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In European conference on computer vision, pages 818–833. Springer, 2014.

[10] Murat Aykanat, O¨ zkan Kılıc¸, Bahar Kurt, and Sevgi Saryal. Classification of lung sounds using convolutional neural networks. EURASIP Journal on Image and Video Processing, 2017(1):1–9, 2017.

[11] Himadri Mukherjee, Sk Md Obaidullah, KC Santosh, Santanu Phadikar, and Kaushik Roy. A lazy learning-based language identification from speech using mfcc-2 features. International Journal of Machine Learning and Cybernetics, 11(1):1–14, 2020.

[12] Jyotibdha Acharya and Arindam Basu. Deep neural network for respiratory sound classification in wearable devices enabled by patient-specific model tuning. IEEE transactions on biomedical circuits and systems, 14(3):535–544, 2020.

[13] Samira Abbasi, Roya Derakhshanfar, Ataollah Abbasi, and Yashar Sarbaz. Classification of normal and abnormal lung sounds using neural network and support vector machines. In 2013 21st Irani an sConference on Electrical Engineering (ICEE), pages 1–4. IEEE, 2013.

[14] R. X. A. Pramono, S. A. Imtiaz and E. Rodriguez-Villegas, "Evaluation of Mel-Frequency Cepstrum for Wheeze Analysis," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* Berlin, Germany, 2019, pp. 4686-4689, doi: 10.1109/EMBC.2019.8857848.

[15] Adam Rao, Emily Huynh, Thomas J Royston, Aaron Kornblith, and Shuvo Roy. Acoustic methods for pulmonary diagnosis. IEEE reviews in biomedical engineering, 12:221–239, 2018.

[16] Mohammed Bahoura and Charles Pelletier. New parameters for respiratory sound classification. In CCECE 2003-Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No. 03CH37436), volume 3, pages 1457–1460. IEEE, 2003.

[17] Fatih Demir, Aras Masood Ismael, and Abdulkadir Sengur. Classification of lung sounds with cnn model using parallel pooling structure. IEEE Access, 8:105376– 105383, 2020.

[18] Yi Ma, Xinzi Xu, and Yongfu Li. Lungrn+ nl: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation. In Interspeech, pages 2902–2906, 2020.

[19] Bruno M Rocha, Dimitris Filos, Lu´ıs Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Niksˇa Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al.
An open access database for the evaluation of respiratory sound classification algorithms. Physiological measurement, 40(3):035001, 2019.

[20] Himadri Mukherjee, Priyanka Sreerama, Ankita Dhar, Sk Obaidullah, Kaushik Roy, Mufti Mahmud, KC Santosh, et al. Automatic lung health screening using respiratory sounds. Journal of Medical Systems, 45(2):1–9, 2021.

[21] ICBHI 2017 challenge.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition (2015). cite. arXiv preprint arxiv:1512.03385.

[23] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.

[24] Gae¨tan Chambres, Pierre Hanna, and Myriam Desainte-Catherine. Automatic detection of the patient with respiratory diseases using lung sound analysis. In 2018 International Conference on Content-Based Multimedia Indexing (CBMI), pages 1–6. IEEE, 2018.

[25] Kirill Kochetov, Evgeny Putin, Maksim Balashov, Andrey Filchenkov, and Anatoly Shalyto. Noise masking recurrent neural network for respiratory sound classification. In
International Conference on Artificial Neural Networks, pages 208–217. Springer, 2018.

[26] Jyotibdha Acharya and Arindam Basu. Deep neural network for respiratory sound classification in wearable devices enabled by patient-specific model tuning. IEEE transactions on biomedical circuits and systems, 14(3):535–544, 2020.

[27] Yi Ma, Xinzi Xu, Qing Yu, Yuhang Zhang, Yongfu Li, Jian Zhao, and Guoxing Wang. Lungbrn: A smart digital stethoscope for detecting respiratory disease using bi- resnet deep learning algorithm. In 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS), pages 1–4. IEEE, 2019.

[29] Music feature extraction in Python - Towards Data Science. (n.d.). Retrieved November 14, 2022, from https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d

[30] Min Xu; et al. (2004). "HMM-based audio keyword generation" (PDF). In Kiyoharu Aizawa; Yuichi Nakamura; Shin'ichi Satoh (eds.). Advances in Multimedia Information Processing – PCM 2004: 5th Pacific Rim Conference on Multimedia. Springer. ISBN 978-3-540-23985-7. Archived from the original (PDF) on 2007-05-10.

[31] What is a spectrogram? Pacific Northwest Seismic Network. (n.d.). Retrieved November 14, 2022, from https://pnsn.org/spectrograms/what-is-a-
spectrogram#:~:text=A%20spectrogram%20is%20a%20visual,energy%20levels%20vary%20over%20 time.

[32] * Chen, C. H., Signal processing handbook, Dekker, New York, 1988

[33] Amoh, J., & Odame, K. (2015). DeepCough: A deep convolutional neural network in a wearable cough detection system. 2015 IEEE Biomedical Circuits and Systems Conference (BioCAS). https://doi.org/10.1109/biocas.2015.7348395

[34] Vinita Shivakumar. Classification of respiratory sounds.

[35] Pedro Sousa Faustino. Crackle and wheeze detection in lung sound signals using convolutional neural networks. 2019.

[36] Xuen Hoong Kok, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. A novel method for automatic identification of respiratory disease from acoustic recordings. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 2589–2592. IEEE, 2019.

[37] Boujelben O, Bahoura M (2018) Efficient FPGA-based architecture of an automatic wheeze detector using a combination of MFCC and SVM algorithms. Journal of Systems Architecture 88:54-64. doi: 10.1016/j.sysarc.2018.05.010

[38] Bardou D, Zhang K, Ahmad S (2018) Lung sounds classification using convolutional neural networks. Artificial Intelligence in Medicine 88:58-69. doi: 10.1016/j.artmed.2018.04.008

[39] Gouyon F., Pachet F., Delerue O. (2000),On the Use of Zero-crossing Rate for an Application of Classification of Percussive Sounds, in Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00 - DAFX-06), Verona, Italy, December 7–9, 2000. Accessed 26 April 2011.

[40] Giannakopoulos, T., & Pikrakis, A. (2014). Audio features. Introduction to Audio Analysis, 59– 103. https://doi.org/10.1016/b978-0-08-099388-1.00004-2

[41] K. El-Maleh, M. Klein, G. Petrucci, P. Kabal, Speech/music discrimination for multimedia applications, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 6, Istanbul, Turkey, IEEE, Piscataway, NJ, June 2000, pp. 2445–2448.

[42] C. Panagiotakis, G. Tziritas, A speech/music discriminator based on RMS and zero-crossings, IEEE Trans. Multimedia 7 (1) (February 2005) 155–166.

[43] B. Kedem, Spectral analysis and discrimination by zero-crossings, IEEE Proc. 74 (1986) 1477–1493.

[44] J. Saunders, Real-time discrimination of broadcast speech/music, in: Proceedings of the IEEE, International Conference on Acoustics, Speech, and Signal Processing, vol. 2, Atlanta, GA, IEEE, Piscataway, NJ, May 1996, pp. 993–996.

[45] C.C. Cheng, C.T. Hsu, Fusion of audio and motion information on hmm-based highlight extraction for baseball games, IEEE Trans. Multimedia 8 (3) (June 2006) 585–599.

[46] M.F. McKinney, J. Breebaart, Features for audio and music classification, in Proceedings of the International Conference on Music Information Retrieval, October 2003.

[47] T. Zhang, Automatic singer identification, in: Proceedings of the IEEE International Conference on Multimedia and Expo, vol. 1, IEEE, Piscataway, NJ, July 2003, pp. 33–36.

[48] Z.J. Chuang, C.H. Wu, Emotion recognition using acoustic features and textual content, in: Proceedings of the IEEE International Conference on Multimedia and Expo, vol. 1, Taipei, Taiwan, IEEE, Piscataway, NJ, June 2004, pp. 53–56.

[49] Support Vector Machine algorithm. GeeksforGeeks. (2022, October 6). Retrieved November 14, 2022, from https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[50] Sinnott, R. O., Duan, H., & Sun, Y. (2016). A case study in Big Data Analytics. *Big Data*, 357– 388. https://doi.org/10.1016/b978-0-12-805394-2.00015-5

[51] L. Wu and L. Li, "Investigating into segmentation methods for diagnosis of respiratory diseases using adventitious respiratory sounds," *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Montreal, QC, Canada, 2020, pp. 768-771, doi: 10.1109/EMBC44109.2020.9175783.

[52] Zumray Dokur. Respiratory sound classification by using an incremental supervised ¨ neural network. Pattern Analysis and Applications, 12(4):309–319, 2009

Source Code:

https://github.com/RafiaAlice/USD-Master-s-Thesis