Department of Electrical and Computer Engineering
North South University

Senior Design Project

Classification of Respiratory Diseases using Machine Learning

1. Name: Md.Jushraf Rahman Chowdhury **ID: 1812901042**

2. Name: Mir Sadia Afrin **ID: 1712366642**

Faculty Advisor

DR. MOHAMMAD ASHRAFUZZAMAN KHAN

Assistant Professor

Department of ECE

Summer 2021

ABSTRACT

Respiratory diseases are leading causes of death and disability in the world. One of the most important indicators of respiratory disorders and health is respiratory sound. There is a direct relation between the sound emitted by a person and the air movement, lung secretions and aerobic and anaerobic changes within the lung. In this research we have tried to automate the diagnosis of respiratory diseases from respiratory sound. We have used the ICBHI 2017 Respiratory Sound Database where there are 920 recordings acquired from 126 individual resulting in a total of 6898 respiration cycles. In this course we have successfully converted the audio files to digital spectrogram and MFCC (mel-frequency cepstrum coefficient) and for the next part of the course we intend to build the Convolutional Neural Network (CNN) model which will label the respiratory diseases. This classification has the potential to detect respiratory abnormalities in early stages of respiratory diseases which can improve the outcome of decision making. It also has the potential to serve the medically underserved population by being used on areas lacking skilled medical staff.

Keywords: *Machine Learning, Convolutional Neural Network (CNN), Respiratory Disease, Respiratory Sound Dataset, Better Accuracy. Better Diagnosis.*

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Respiratory diseases, or lung diseases, are pathological conditions affecting the organs and tissues that make gas exchange difficult in air-breathing animals. The symptoms can be depended on the respiratory condition, such as - shortness of breath with activity, persistent cough, wet cough with mucus (COPD, cystic fibrosis, asthma), dry, "nonproductive" cough (pulmonary fibrosis), chest tightness, wheezing, rapid shallow breathing, hoarseness. It causes various diseases like lung cancer, asthma, Tuberculosis, Chronic Obstructive Pulmonary Disease (COPD), etc. Every year, a lot of people die from respiratory diseases. Still in Bangladesh, the lacking of immediate diagnosis of the diseases increasing the mortality rate.

Our goal is to detect respiratory anomalies in early stages which can enhance the diagnosis process and play a role in decreasing respiratory fatalities.

1.1 WHO DATA

In Bangladesh, respiratory diseases have been increasing rapidly. According to WHO, every year thousands of people die from various respiratory diseases. The latest WHO data published in 2018 -

- Lung Cancers Deaths in Bangladesh reached 13,265 or 1.71% of total deaths.
- Tuberculosis Deaths in Bangladesh reached 65,746 or 8.47% of total deaths.
- Asthma Deaths in Bangladesh reached 14,674 or 1.89% of total deaths.

1.2 LEADING CAUSES OF DEATH IN BANGLADESH

Respiratory diseases are on lead caused death in Bangladesh. And many of these happen because of misdiagnosis, late diagnosis, unserved treatment etc. Here are some lead diseases which caused deaths in Bangladesh:

- **Lung Cancer**

  Lung Cancer is one of the most common disease in Bangladesh which causes death

every year and according to newest study, the cases are rising rapidly. *According to the latest Hospital Cancer Registry Study, there is a reportedly near 200% rise in the country's lung cancer burden in just three years.*

- **Lower Respiratory Infections**

  About 7% of all death in Bangladesh are caused by lower respiratory infections and it is leading second in mortality rate in the country. 28% of children death are also caused by various respiratory infections.

- **Chronic Obstructive Pulmonary Disease**

  Chronic Obstructive Pulmonary Disease (COPD) is one of the dangerous diseases that caused death in Bangladesh very often. This is a collective term for ailments which affect the lung. These cause emphysema and bronchitis. COPD causes 7% of all deaths in this country.

- **Tuberculosis**

  According to World Health Organization, Bangladesh is the "High TB Burden Country" because of tuberculosis. 3% of all deaths in Bangladesh are caused by tuberculosis.

- **Asthma**

  Asthma is one of the most common disease that every age of people can have in Bangladesh. Even there is a huge record of children who has asthma at a very early age.

- **COVID-19**

  COVID-19 is now the most caused disease witch mainly affecting the lung and destroy the whole respiratory system and causing enormous deaths every day not just in Bangladesh, but also in the world.

1.3 AIM

We aim to automate the diagnosis of respiratory diseases in early stages by classifying them from breathing recordings by the help of using machine learning. We will use the sound data sample and create a machine learning model which will give a better accuracy than existing architectures.

CHAPTER 2: LITERATURE REVIEW

While working on this research project, we studied various research papers related to our project topic and picked a few papers from there which were based on predicting Respiratory Diseases with better accuracy. We chose these particular papers because their working approach is closely related to our research work. We also deduced some ideas from there for doing our work.

2.1EXISTING LITERATURE EXPLANATION

These following research works are based on different types of respiratory diseases to predict them in various ways. Scientists have been studying these for years to improve the prediction accuracy.

2.1.1. Lung Disease Classification using Deep Convolutional Neural Network

This is a study by "Zeenat Tariq, Sayed Khushal Shah, Yugyung Lee" from *"School of Computing and Engineering, University of Missouri-Kansas City, USA"* in November 2019. While modern technologies are improving on medicine, researchers are working on more precise and customized treatment strategies for various diseases. This paper aimed for making a better accuracy on the medical field by using deep learning with available datasets. By extracting spectrogram features and labels of the lung sounds sample data to 2D Convolutional Neural Network (CNN) model. Then, normalizing the sound data by removing the peak values and noise, the classifications are made using the deep learning process.

The final accuracy has been gained approximately 97%. This model helps to get an accuracy which will ben accepted in the medical field for better performance.

2.1.2. Asthma and Chronic Obstructive Pulmonary Disease (COPD) – Differences and Similarities

This is a study by "Vesna Cukic, Vladimir Lovre, Dejan Dragisic, Aida Ustamujic" from *"Clinic for Pulmonary Diseases and TB "Podhrastovi", Clinical center of Sarajevo University, Bosnia and Herzegovina"* in 2012.

This paper is focused on Asthma and Chronic Obstructive Pulmonary Disease (COPD). These diseases are highly common in todays world, affecting millions of people. Where Asthma is very common and serious problem, COPD has been causing serious damages like chronic morbidity and mortality. As there are many similarities between Asthma and COPD, there are many differences too such as, differences in etiology, symptoms, type of airway inflammation, inflammatory cells, mediators, consequences of inflammation, response to therapy, course.

So, the aim is to discuss the similarities and differences of these diseases to better understanding pf pathology to do the correct diagnosis for better treatment to reduce the morbidity and mortality.

2.1.3. LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm

This is a conference paper by "Yi Ma, Xinzi Xu, Qing Yu, Yuhang Zhang, Yongfu Li*, Jian Zhao and Guoxing Wang" from *"Department of Micro-Nano Electronics and MoE Key Lab of Artificial Intelligence Shanghai Jiao Tong University, Shanghai, China"* in Octiober 2019.

To improve health care services for immediate diagnosis, this study is to help the medical staff. The idea is of a digital stethoscope to predict the disease with better accuracy from the basic lung sound to have the immediate diagnosis for respiratory related problems.

The idea is to develop an improved bi-ResNet deep learning architecture, LungBRN, which uses STFT and wavelet feature extraction techniques to improve the accuracy.

This technique has achieved a performance of 50.16% accuracy. Compared to all participating teams from ICBHI 2019.

2.1.4. A Respiratory Sound Database for the Development of Automated Classification

This is a conference paper by "B. M. Rocha[1], D. Filos[2], L. Mendes[1], I. Vogiatzis[2], E. Perantoni[2], E. Kaimakamis[2], P. Natsiavas[2], A. Oliveira[3,4] , C. Jácome[3], A. Marques[3,4], R. P. Paiva[1], I. Chouvarda[2], P. Carvalho[1], N. Maglaveras[2]" from *" [1] Centre for Informatics and Systems, Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra, Coimbra, Portugal; [2] Lab of Computing, Medical Informatics and Biomedical Imaging Technologies, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece; [3] Lab3R - Respiratory Research and Rehabilitation Laboratory, School of Health Sciences, University of Aveiro, Aveiro, Portugal; [4] Institute for Biomedicine (iBiMED), University de Aveiro, Portugal "* in November 2017.

This paper has focused on the development of the automatic analysis of respiratory diseases by respiratory sound database. As the publicly available large datasets have been used in various new algorithms to evaluate, still there is lacking of better result.

So, in this paper, the idea is to collect respiratory sound data and use it to develop an automated classification of the diseases.

2.1.5. XRAY AI: Lung Disease Prediction Using Machine Learning

This is an International Journal of Information Systems and Computer Sciences by "Justin Monsi[1], Justine Saji[2], Keerthy Vinod[3], Liya Joy[4], Jis Joe Mathew[5]" from *" 1 Amal Jyothi*

College of Engineering, Kottayam, India; 2 Amal Jyothi College of Engineering, Kottayam, India; 3 Amal Jyothi College of Engineering, Kottayam, India; 4Amal Jyothi College of Engineering, Kottayam, India; 5 Amal Jyothi College of Engineering, Kottayam, India " in March-April 2019.

In this paper, the NIH chest X-ray dataset comprised of 112,120X-ray images with disease labels from 30,000 unique patients has been presented. The accuracy is expected to achieve more than 90%. The deep convolutional neural network has been used to diagnosis the disease patterns and locations. Also, the automated detection system is being used.

The goal is to create a system to get better accuracy and do the correct diagnosis without getting a second opinion by reducing misdiagnosis and fatal consequences.

2.1.6. Improving Disease Prediction by Machine Learning

This is an International Research Journal of Engineering and Technology (IRJET) by "Smriti Mukesh Singh[1], Dr. Dinesh B. Hanchate[2]" from *"1 Student of Computer Engineering, Pune University, VPKBIET, Baramati, India; 2 Professor of Computer Engineering, Pune University, VPKBIET, Baramati, India"* in June 2018.

As now biomedical information has a huge data collection. From that big data, medical services need to investigate medicinal information, diagnosis different types of diseases and also maintain the administration. All of these become very difficult with the big amount of data with a proper model, especially disease prediction. As the importance of diagnosis is very high, this paper discussed how to improve the disease prediction using machine learning. The goal of this paper is to improve the disease prediction with better accuracy and correct diagnosis using machine learning model.

CHAPTER 3: METHODOLOGY

This chapter provides a chronological outline of the different sections of the work. We will talk about the theories, techniques and step by step workflow of our work.

WORKFLOW:

A rudimentary diagram of our work method is shown below:

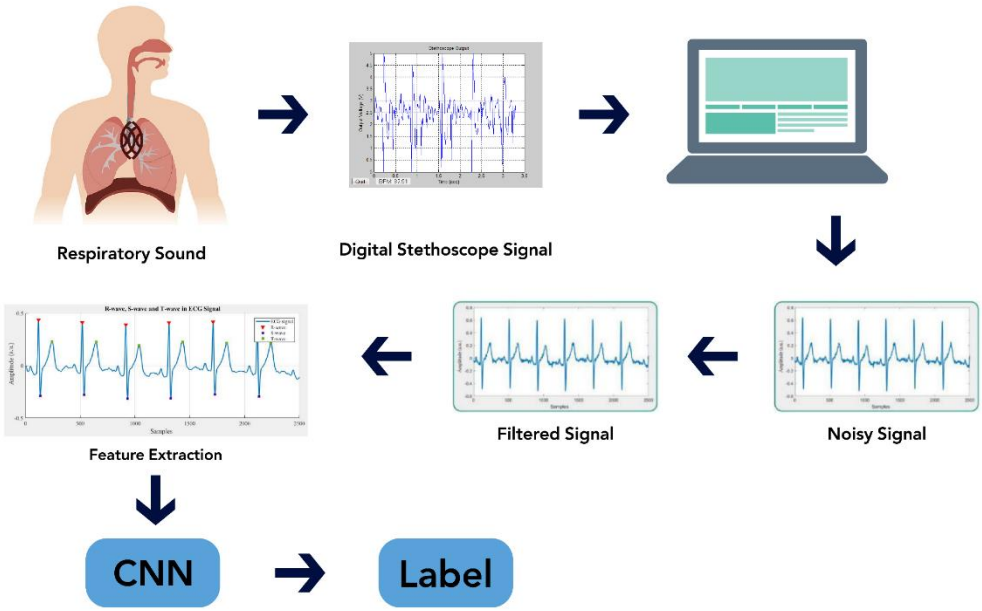

Figure: Fundamental diagram of our work

The above diagram portrays the overall steps of our work for a better understanding. We will record respiratory or breathing sounds of individuals through a digital stethoscope or the listed recording equipment. From there the audio files (.wav format) will stored in a computer and the work of our model starts from there. The noise from the audio files will be filtered out and then will be converted to digital spectrogram and MFCC (mel-frequency cepstrum coefficient) after which it will be passed on to our CNN model. Our model will finally put a label as to which category of the respiratory diseases a certain audio file belonged to. In this way we will classify audio files into different respiratory diseases.

DATASET:

We will use the ICBHI 2017 Respiratory Sound Database. This was created by two research teams in Portugal and Greece. It includes 920 annotated recordings of varying length - 10s to 90s. These recordings were taken from 126 patients. There are a total of 5.5 hours of recordings containing 6898 respiratory cycles - 1864 contain crackles, 886 contain wheezes and 506 contain both crackles and wheezes. The data includes both clean respiratory sounds as well as noisy recordings that simulate real life conditions. The patients span all age groups - children, adults and the elderly.

This dataset includes:

- 920 .wav sound files
- 920 annotation .txt files
- A text file listing the diagnosis for each patient
- A text file explaining the file naming format
- A text file listing 91 names (filename_differences.txt)
- A text file containing demographic information for each patient

PREPROCESSING:

Since our dataset is of audio type, it is a complicated process for the computer to understand it. So our main focus on this course (499A) has been to convert the audio file to a visual file that is an image. We were successful in converting the audio files to both digital spectrogram and MFCC (mel-frequency cepstrum coefficient).

FEATURE EXTRACTION:

So far the features that we extracted from the audio files through our digital spectrogram and MFCC are:

Feature Name	Description
Frequency	From the digital spectrogram we were able to see the intensity of sound at different frequencies
Pitch	From the MFCC we were able to figure out the nature of pitch of the audio file. Higher than origin meant high pitch sound and lower meant low pitch.

MACHINE LEARNING MODELS:

We will be using the following model:

CNN: A convolutional neural network (CNN) is a type of artificial neural network used in image recognition and processing that is specifically designed to process pixel data. CNNs are powerful image processing, artificial intelligence (AI) that use deep learning to perform both generative and descriptive tasks, often using machine vison that includes image and video recognition, along with recommender systems and natural language processing.

CHAPTER 4: RESULTS

Our main focus on this course was to preprocess the dataset where the primary goal was to convert the audio files to visual images. We converted some of the audio files to digital spectrogram and MFCC where we found the following results:

Digital Spectrogram:

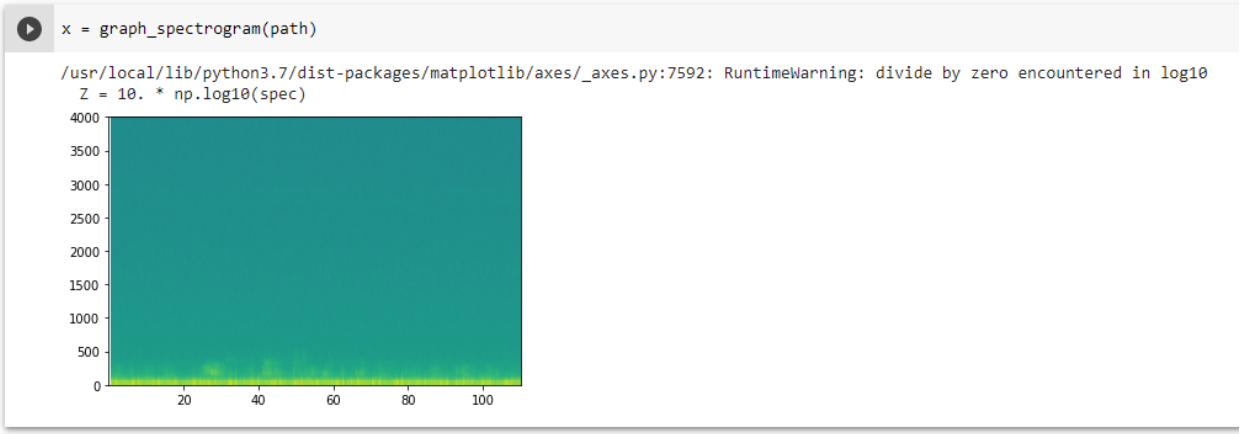

Figure: Conversion of sound to digital spectrogram

As we can see, the loudness of the sound can be identified through this image. At x-axis there is time and on y-axis there is frequency. At different time we can see the frequency of the sound and can judge its intensity or loudness.

MFCC:

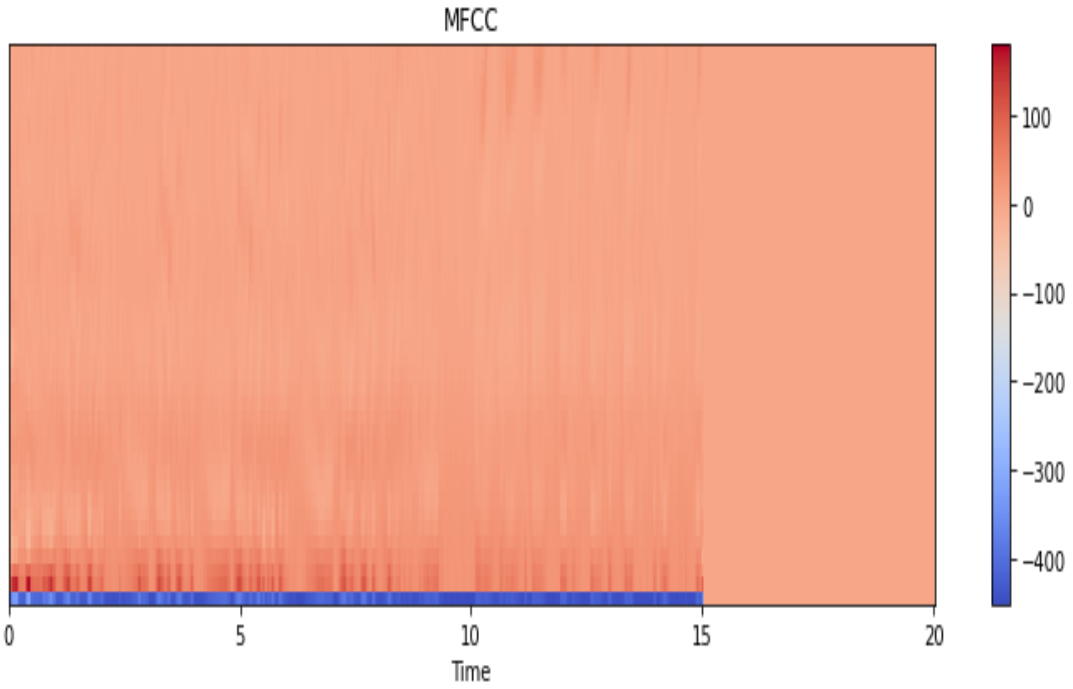

Figure: Conversion of sound to MFCC

From the MFCC we can judge the quality of pitch. The higher the y-coordinates are from origin the higher is the pitch and the lower it is from the origin the lower is the pitch.

CHAPTER 5: CONCLUSION

In this chapter, we are going to discuss how our work can serve as a benchmark in the medical industry and pave path for future significant works.

5.1 DISCUSSION

In this course our significant achievement was to convert the audio file to a visual image. We achieved that by converting audio files (.wav format) to a digital spectrogram and MFCC (mel-frequency cepstrum coefficient). We want to preprocess the data further to make it suitable for sending to the CNN model which we will build on the next part of the course. We also plan on broadening the dataset to dive deep into the causes and hope for better findings.

5.2 SUMMARY

The climate is deteriorating faster than ever. The air quality that we have today has degraded many times compared to the air quality we had 10 years earlier. As a result the number of deaths and sickness due to respiratory disease have increased manifold. But unfortunately this is an area very under looked upon as there are very few papers and works related to serve this sector. Our work can play a significant role in this sector. There are a lot of areas in the world where there is huge lacking of skilled medical staff. Highly experienced doctors are very expensive and not everyone can afford them. Also many doctors hesitate to serve in remote areas. As a result the people living in those areas are medically underserved. In such circumstances just a recording though a digital stethoscope and run-through via our model can provide immediate diagnostic reports. It can tell what type of respiratory disease that person has. It does not require any skilled medical staff to record audio and to pass it into our model.

5.3 FUTURE WORK

In future, we would also like to add more data to test our algorithm to make it more credible for real-world use. In the next course 499B we are focusing on creating a CNN model which will successfully classify our audio files and label what disease that audio file belonged to.

REFERENCES

[1] Zeenat Tariq, Sayed Khushal Shah, Yugyung Lee, "Lung Disease Classification using Deep Convolutional Neural Network" *IEEE International Conference on Bioinformatics and Bio-medicine 2019, At: San Diego, California*, Nov. 2019. [online]. Available at: (PDF) Lung Disease Classification using Deep Convolutional Neural Network (researchgate.net)

[2] Vesna Cukic, Vladimir Lovre, Dejan Dragisic, and Aida Ustamujic,"Asthma and Chronic Obstructive Pulmonary Disease (COPD) – Differences and Similarities", v.24(2); 2012. [online]. Available at: Asthma and Chronic Obstructive Pulmonary Disease (COPD) – Differences and Similarities (nih.gov)

[3] 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", *BioCAS*, Oct. 2019. [online]. Available at: (PDF) LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm (researchgate.net)

[4] B. M. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, R. P. Paiva, I. Chouvarda, P. Carvalho, N. Maglaveras, "A Respiratory Sound Database for the Development of Automated Classification" *ICBHI 2017*, Nov. 2017. [online]. Available at: (PDF) A Respiratory Sound Database for the Development of Automated Classification (researchgate.net) [Accessed Nov- 2017].

[5] Justin Monsi, Justine Saji, Keerthy Vinod, Liya Joy, Jis Joe Mathew, "XRAY AI: Lung Disease Prediction Using Machine Learning", Volume 8, No.2, March - April 2019. [online]. Available at: http://warse.org/IJISCS/static/pdf/file/ijiscs12822019.pdf

[6] Smriti Mukesh Singh, Dr. Dinesh B. Hanchate, “Improving Disease Prediction by Machine Learning”, Volume: 05 Issue: 06 | June-2018. [online]. Available at: (PDF) Improving Disease Prediction by Machine Learning (researchgate.net)

[7] Rohan Tan Bhowmik, “A Multi-Modal Respiratory Disease Exacerbation Prediction Technique Based on a Novel Spatio-Temporal Machine Learning Architecture and Local Environmental Sensor Networks” *et al. · The Harker School*, Apr. 3, 2021. [online]. Available at: 2103.03086.pdf (arxiv.org)