Lung Sound Classification using Light-Weight Architectures

A PROJECT REPORT
SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
Master of Technology
IN
Faculty of Engineering

BY

Vasu Mehra

Computer Science and Automation
Indian Institute of Science
Bangalore – 560 012 (INDIA)

June, 2024

DEDICATED TO

My Parents and Siblings

without whom I wouldn't be here

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to my advisors Prof. Siddharth Barman and Prof. Prasanta Kumar Ghosh for their continuous support, guidance and constructive feedback and criticism throughout the course of this project. Repeated discussions with them gave profound insights and shaped the ideas contributing to the completion of this project. I would also like to thank my advisors from AI Health Highway Pvt. Ltd., Dr (Maj) Satish S Jeevannavar, Dr. Alex Paul Kamson and Akshay Sawant for their continued support in terms of keeping me up with current research in the field of lung sound analytics, and providing me with continued support in terms of code quality. I would also like to thank the Department of Computer Science and Automation, IISc for providing me with the opportunity to work on this project and the continuous availability of GPUs and Linux servers for data storage and processing.

Abstract

India has one of the highest rates of chronic respiratory disorders, according to the most recent Global Burden of Disorders Report, 2017. Although 15.69% of all chronic respiratory diseases worldwide are caused by India, the country accounts for 30.28% of all chronic respiratory disease-related fatalities worldwide. With a staggering 55.23 million cases of Chronic Obstructive Pulmonary Disease, India leads the globe in this condition! India accounts for almost 0.85 million of the world's Chronic Obstructive Pulmonary Disease-related fatalities, which is the second-highest figure worldwide. India also tops the globe in asthma-related mortality, accounting for 43% of all asthma-related deaths worldwide. This project uses the technological developments in machine learning and artificial intelligence combined with the knowledge of signal processing to identify patterns in different adventitious sounds like crackles and wheezes. The power of convolution has been harnessed by using efficient feature extraction techniques like mel-spectrogram and mel-frequency cepstral coefficients. Later, we also used transformer architecture to extract even more useful representations of these different lung sounds, which helped us detect any adventitious sound or identify the disease associated with the sound recording. We compare our results for lightweight CNN architecture, the original Wav2Vec2 model developed by Facebook, the Wav2Vec2 model pre-trained on heart and lung sounds, and a lighter parameter version of the same Wav2Vec2 model to make an assessment of the trade-off between number of parameters and model-accuracy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Lung diseases pose a significant public health challenge globally, with their prevalence and impact particularly pronounced in countries like India. Recent health surveys underscore the grave consequences of these ailments, shedding light on their far-reaching effects on individuals, communities, and healthcare systems. In this introduction, we elucidate the adverse effects of lung diseases in India, drawing upon empirical evidence from contemporary health studies. Subsequently, we delve into the potential of deep learning architectures to mitigate this burgeoning health crisis.

India, home to over 1.3 billion people, grapples with many health issues, among which respiratory diseases stand out as a formidable concern. According to the Global Burden of Disease Study 2019[1], chronic respiratory diseases accounted for a staggering 115.9 million disability-adjusted life years (DALYs) in India, signifying the profound impact of these conditions on morbidity and mortality rates. Furthermore, the burden of lung diseases extends beyond mere statistics, permeating various facets of societal well-being.

Recent health surveys conducted across different regions of India have furnished alarming insights into the prevalence and consequences of lung diseases. The National Family Health Survey (NFHS-5), from 2019 to 2021, revealed a concerning rise in respiratory ailments, with significant proportions of the population reporting symptoms indicative of conditions such as chronic obstructive pulmonary disease (COPD), asthma, and pneumonia. These findings underscore the pervasive nature of lung diseases and the urgent need for effective interventions to alleviate their impact.

The ramifications of lung diseases reverberate across socioeconomic strata, exacerbating health inequities and impeding economic development. In India, where a substantial portion of the population grapples with poverty and inadequate access to healthcare services, the burden of respiratory ailments is disproportionately borne by vulnerable communities. Moreover, the in-

direct costs associated with lost productivity and healthcare expenditures further strain already overburdened healthcare systems, perpetuating a vicious cycle of illness and impoverishment.

Recent advancements in deep learning have catalyzed innovative approaches for analyzing lung sounds, offering novel insights into diagnosing and monitoring respiratory conditions. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory Networks (LSTMs) have emerged as indispensable tools in this endeavour, each bringing unique capabilities to the forefront of lung sound analysis.

CNNs, renowned for their proficiency in image recognition tasks, have been adapted to spectrogram representations of lung sounds, facilitating the extraction of discriminative features from audio data. By convolving over spectrograms with learned filters, CNNs can identify salient patterns indicative of specific respiratory pathologies, such as wheezes, crackles, and airflow obstruction. Studies such as that by Rajpurkar et al. (2017) have demonstrated the efficacy of CNNs in automating the detection of abnormal lung sounds, showcasing their potential for enhancing diagnostic accuracy and efficiency.

Meanwhile, RNNs and LSTMs have garnered attention for their ability to capture temporal dependencies in sequential data, making them well-suited for analyzing the dynamic nature of lung sounds over time. By processing audio signals as sequential inputs, RNN-based architectures can discern nuanced patterns and temporal relationships within lung sound recordings, enabling finer-grained analysis of respiratory dynamics. Notably, research by Gulrajani et al. (2020) has leveraged RNNs to detect and classify pathological events in respiratory sounds, showcasing the utility of recurrent models in real-time monitoring and disease surveillance.

Furthermore, integrating CNNs with RNNs in hybrid architectures has yielded synergistic benefits, enabling comprehensive analysis of spectral and temporal features in lung sound data. By combining convolutional filters with recurrent connections, these hybrid models can leverage the strengths of both architectures, thereby enhancing the robustness and interpretability of lung sound analysis. Notable examples include the work by Attia et al. (2019), which proposed a CNN-RNN hybrid for automated detection of abnormal breath sounds, demonstrating superior performance compared to individual models.

In summary, CNNs, RNNs, and LSTMs represent versatile tools for lung sound analysis, offering complementary capabilities for extracting meaningful insights from audio data. As evidenced by the growing body of literature, deep learning-based approaches hold immense promise for revolutionizing respiratory healthcare, paving the way for more accurate diagnosis, proactive monitoring, and personalized interventions in managing lung diseases.

Chapter 2

Understanding the audio signal

Each audio signal, by nature, has different properties. They have different fundamental frequency ranges and, as a result of it, different harmonics. Hence, it is crucial to understand the audio signals we will work with and the properties of such sound signals.

2.1 Lung Exam

A doctor conducts a lung examination by following a systematic approach to assess the respiratory system for abnormalities. The examination typically begins with a thorough patient history, focusing on symptoms such as cough, shortness of breath, chest pain, and any history of smoking or respiratory diseases. The physical examination itself starts with inspection, where the doctor observes the patient's chest for asymmetry, deformities, or abnormal movements. This is followed by palpation, where the doctor uses their hands to feel for any unusual masses, tenderness, or vibrations (fremitus) on the chest wall.

Next, the doctor performs percussion, tapping on the chest wall with fingers to assess the underlying structures. This helps determine if the lungs are filled with air, fluid, or solid material. Normal lungs produce a resonant sound, whereas dullness may indicate fluid or a mass, and hyperresonance could suggest hyperinflation as seen in conditions like emphysema.

Auscultation is a crucial step, where the doctor uses a stethoscope to listen to lung sounds. The patient is asked to breathe deeply through the mouth. The doctor listens for normal breath sounds, such as vesicular sounds, and identifies any abnormal sounds like wheezes, crackles, or rhonchi. Wheezes may indicate airway obstruction, while crackles could suggest fluid in the airways, often associated with conditions like pneumonia or heart failure. The ausculation points are shown in Figure 2.1.

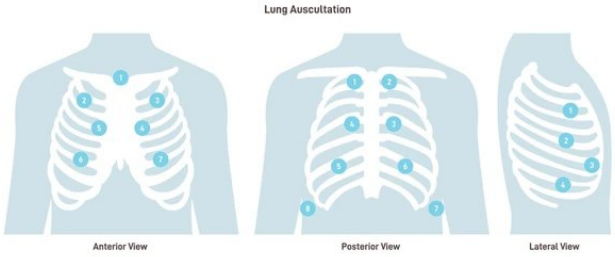


Figure 2.1: Auscultation points on the anterior, posterior and lateral parts of chest

The examination may also include specific maneuvers, such as asking the patient to say "ninety-nine" to detect vocal resonance changes, or having the patient cough to see if certain sounds clear. These findings are integrated with clinical context and diagnostic tests like chest X-rays or pulmonary function tests to form a comprehensive assessment. This methodical approach ensures a thorough evaluation of the respiratory system, aiding in accurate diagnosis and effective treatment planning.

2.2 Normal Lung Sound

Healthy lung sounds, known as vesicular breath sounds, have distinct spectral and auditory properties when auscultated through a stethoscope. These sounds are typically low-pitched, with frequencies ranging from about 100 to 600 Hz, and are soft and rustling in nature. The sound is generated by the turbulent airflow in the small airways and alveoli during inhalation. Spectrally, vesicular sounds exhibit a higher amplitude during inspiration, diminishing during expiration due to the passive nature of exhalation in normal breathing. This sound lacks prominent high-frequency components or distinct harmonics, indicating a smooth and unobstructed airflow.

Auditory characteristics of healthy lung sounds include a smooth, uninterrupted airflow, with no adventitious sounds such as crackles, wheezes, or rhonchi, which are indicative of pathologies. The intensity of vesicular sounds is generally louder at the lung bases due to greater ventilation-perfusion matching and decreased over areas with muscle or bone coverage. During inspiration, the sound has a soft crescendo, peaking at the end of inhalation, followed by a quieter expiratory phase. This auditory pattern helps clinicians distinguish normal lung function from abnormal conditions. The clear, rhythmic pattern of vesicular sounds reflects unobstructed airways and healthy alveolar function, crucial for effective gas exchange in the lungs. Regular auscultation helps in early detection of respiratory anomalies, ensuring timely medical intervention. The spectrogram representation can be found in Figure 2.2.

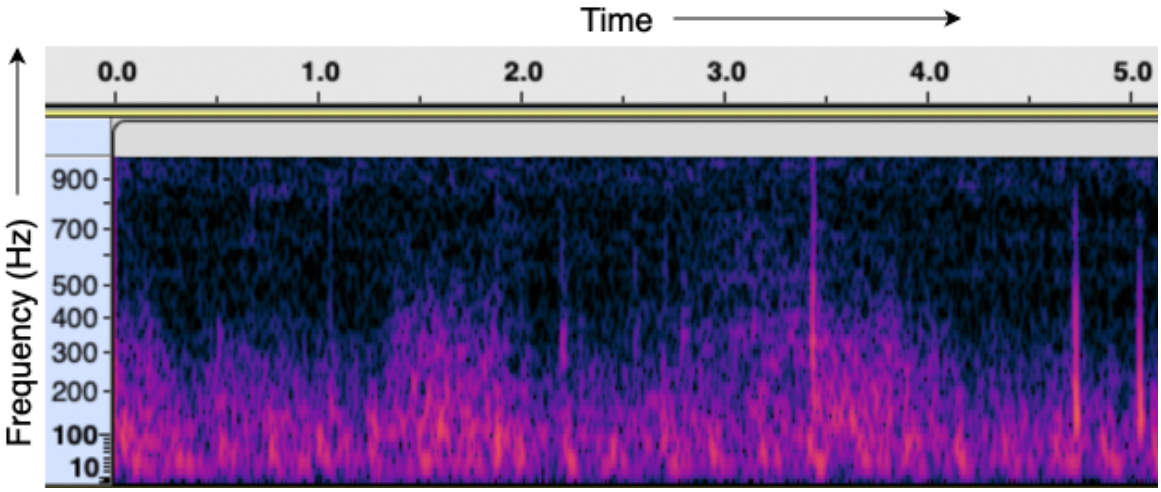


Figure 2.2: Spectrogram representation of a normal lung sound

2.3 Crackles

Crackles, also known as rales, are abnormal lung sounds characterized by their distinctive spectral and auditory properties when heard through a stethoscope. Spectrally, crackles are typically high-pitched and can range from 400 to 2000 Hz. They are brief, discontinuous, and occur in rapid succession, creating a crackling or popping sound. Crackles are further categorized into fine and coarse crackles based on their pitch, duration, and sound quality. Fine crackles are high-pitched, short in duration, and occur during the late phase of inspiration, often indicating conditions such as interstitial fibrosis or early pulmonary edema. Coarse crackles, on the other hand, are lower-pitched, longer in duration, and can be heard during both inspiration and expiration, commonly associated with chronic bronchitis, bronchiectasis, or resolving pneumonia.

Auditorily, crackles resemble the sound of Velcro being pulled apart or the crackling of a fireplace. Fine crackles are subtle and soft, requiring careful auscultation, while coarse crackles are louder and more pronounced. The presence of crackles is indicative of fluid or secretions in the small airways and alveoli, caused by conditions that disrupt the normal air-liquid interface in the lungs. This disruption leads to the sudden opening of collapsed airways or the movement of air through fluid, producing the characteristic crackling sound.

The timing of crackles in the respiratory cycle is clinically significant; late inspiratory crackles often suggest interstitial lung disease or congestive heart failure, while early inspiratory or expiratory crackles are more indicative of obstructive airway diseases. Understanding the spectral and auditory properties of crackles aids clinicians in diagnosing and differentiating between various pulmonary conditions, guiding appropriate treatment and management strategies. The

spectrogram representation can be found in Figure 2.3.

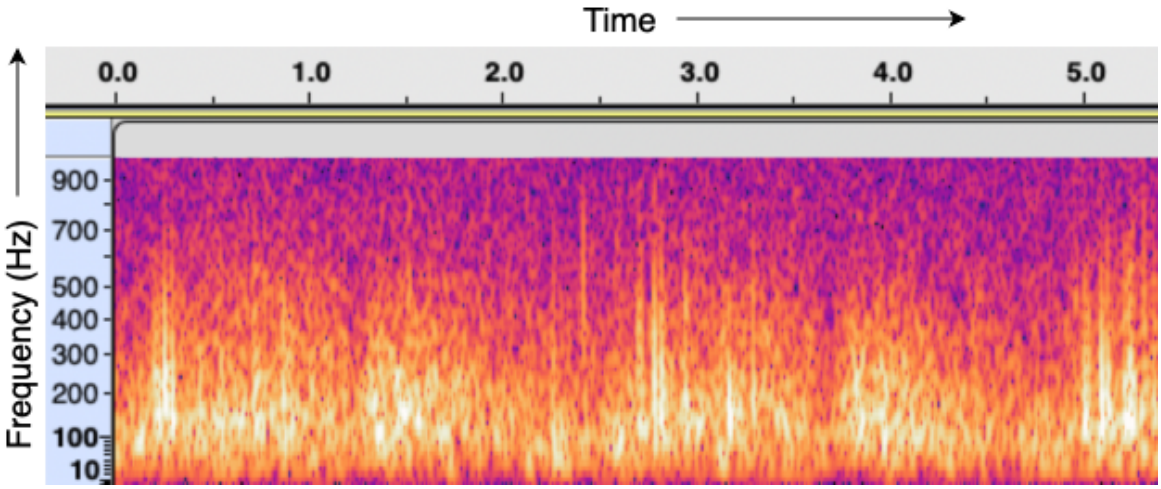


Figure 2.3: Spectrogram representation of a Crackle in lung sound

2.4 Wheezes

Wheezes are abnormal lung sounds that exhibit distinct spectral and auditory properties when auscultated through a stethoscope. Spectrally, wheezes are characterized by their high-pitched, continuous nature, with frequencies ranging from 400 Hz to over 2000 Hz. They are musical sounds that arise from oscillations in the airway walls when airflow is obstructed or turbulent. The pitch of the wheeze depends on the degree of airway narrowing and the velocity of airflow; tighter constrictions and higher airflow velocities produce higher-pitched wheezes.

Auditorily, wheezes are typically described as squeaky or whistling sounds. They can be heard during both inspiration and expiration, but are often more prominent during expiration due to the increased pressure in the chest cavity, which exacerbates airway narrowing. Wheezes can vary in duration and intensity, and may be polyphonic (multiple tones) or monophonic (single tone), depending on whether they arise from multiple bronchial tubes or a single site of obstruction, respectively.

The presence of wheezing is often associated with conditions that cause airway obstruction, such as asthma, chronic obstructive pulmonary disease (COPD), bronchitis, and reactive airway disease. In asthma, wheezing is typically diffuse and polyphonic, reflecting widespread bronchoconstriction. In contrast, monophonic wheezes suggest a localized obstruction, such as a tumor or foreign body.

Clinically, the detection and analysis of wheezes help in diagnosing the severity and location of airway obstructions. Continuous wheezing suggests persistent airway narrowing, while the

improvement or disappearance of wheezing can indicate a response to treatment. Recognizing the spectral and auditory properties of wheezes is essential for effective respiratory assessment and management. The spectrogram representation can be found in Figure 2.4.

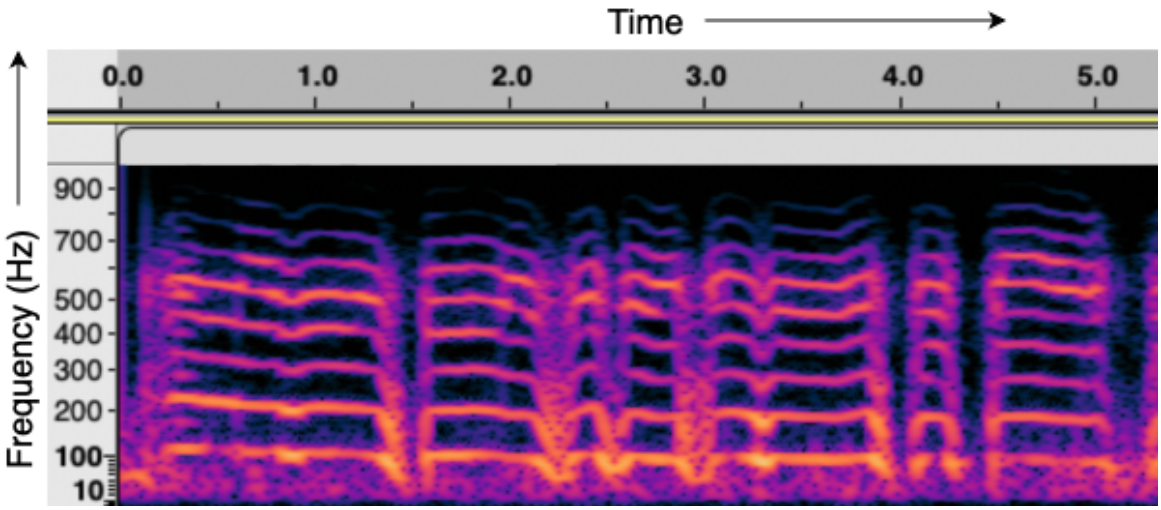


Figure 2.4: Spectrogram representation of a Wheeze in lung sound

Chapter 3

Literature Review

3.1 Crackle and wheeze detection in lung sound signals using convolutional neural networks, Faustino, et. al.

Faustino et al.[3] explore the use of convolutional neural networks (CNNs) for detecting crackles and wheezes in lung sound recordings. These abnormal breath sounds can be crucial indicators of respiratory illnesses. The proposed system utilizes mel spectrograms as input features for the CNN. Mel spectrograms convert raw audio data into a visual representation highlighting the distribution of sound energy across frequencies. The CNN architecture is likely to include convolutional layers for extracting relevant patterns from the mel spectrogram and fully connected layers for final classification. The authors evaluate different feature extraction methods (raw audio, power spectral density, mel spectrogram, MFCCs) and find mel spectrograms to yield the best performance. While achieving accuracy and sensitivity comparable to existing methods (around 43% and 51% respectively), the results suggest room for improvement. Limitations include the relatively low accuracy and the potential for further optimization of the CNN architecture or training process.

3.2 Different Transfer Learning Approaches for Recognition of Lung Sounds: Review, Jemisha, et. al.

The review by Jemisha and Milin[8] focuses on transfer learning approaches for lung sound classification. Transfer learning leverages pre-trained models on vast image datasets for lung sound analysis. The paper explores using pre-trained models like VGG16, ResNet, and AlexNet

as feature extractors. These models, originally designed for image recognition, are repurposed for lung sounds by converting them into visual representations like mel spectrograms. The pre-trained models then act as feature extractors, identifying patterns within the mel spectrograms. A separate classifier is then trained on these extracted features to categorize lung sounds as normal, wheezes, crackles, or a combination. This approach leverages the extensive knowledge learned by the pre-trained models from image datasets, reducing the need for large lung sound specific datasets for training. The review highlights studies comparing these transfer learning approaches with traditional deep learning models, finding that CNNs with transfer learning often achieve superior performance. However, limitations include the "black-box" nature of deep learning models, making it difficult to understand how they arrive at their classifications. Additionally, the reliance on pre-trained models on image data introduces an inherent mismatch when applied to audio data like lung sounds.

3.3 Lung Sound Classification Using Snapshot Ensemble of Convolutional Neural Networks, Nguyen et. al.

Nguyen and Pernkopf[7] propose a lung sound classification system utilizing a "snapshot ensemble" of convolutional neural networks (CNNs). This approach aims to improve robustness and efficiency in classifying lung sounds. Each CNN within the ensemble is trained on a "snapshot" of the overall training data, providing a diverse set of models. During testing, the individual predictions from each CNN are combined using majority voting to deliver the final classification.

The paper demonstrates that the snapshot ensemble approach outperforms single CNN models, achieving higher accuracy in classifying normal lung sounds from abnormal ones. This suggests the ensemble mitigates overfitting and leverages the strengths of various models trained on different data subsets. However, the computational cost associated with training and deploying an ensemble of models is a limitation to this approach.

3.4 Self-Supervised Audio Encoder with Contrastive Pre-training for Respiratory Anomaly Detection, Kulkarni et. al.

The paper "Self-Supervised Audio Encoder with Contrastive Pretraining for Respiratory Anomaly Detection"[6] presents a novel approach to detect respiratory anomalies using self-supervised learning methods. Authored by Shubham Kulkarni, Hideaki Watanabe, and Fuminori Homma, this research was presented at the IEEE International Conference on Acoustics, Speech, and

Signal Processing (ICASSP) 2023. The study introduces a self-supervised audio encoder that leverages contrastive pretraining to improve the detection of respiratory anomalies from lung sound recordings. This method involves a two-stage training process. Initially, the model undergoes self-supervised pretraining using a large acoustic dataset to extract high-level audio representations. This is followed by fine-tuning on a respiratory sound dataset to specialize the model for anomaly detection tasks.

The encoder's architecture is designed to handle direct waveform inputs, which allows it to capture intricate features of respiratory sounds. The contrastive learning technique used in pretraining helps the model differentiate between normal and abnormal lung sounds more effectively by maximizing the similarity between augmented versions of the same audio sample while minimizing the similarity between different samples. This approach addresses the challenge of limited labeled data by utilizing large amounts of unlabeled audio data for pretraining, thereby improving the model's robustness and accuracy in identifying respiratory anomalies. The paper highlights significant improvements in anomaly detection performance compared to traditional supervised learning methods, showcasing the potential of self-supervised learning in medical audio analysis.

Chapter 4

Dataset

In this chapter, we explore the various open source datasets that have been used for our lung sound analysis for adventitious sound detection and lung disease detection.

4.1 ICBHI Dataset

The ICBHI (International Conference on Biomedical Health Informatics) Lung Sound Dataset[9] is a comprehensive collection of respiratory sounds designed to aid in the development and evaluation of algorithms for lung sound analysis. This dataset was compiled for the 2017 ICBHI challenge, focusing on the automatic classification of respiratory sounds. It includes recordings from 126 subjects, encompassing both healthy individuals and patients with various respiratory conditions such as asthma, COPD, and lower respiratory tract infections. The dataset features 6,898 audio samples, captured using both digital stethoscopes and smartphones. These recordings are meticulously labeled with annotations specifying the type of respiratory sound present, such as normal breath sounds, wheezes, crackles, or both. Additionally, the dataset includes detailed metadata, including patient information like age, gender, and diagnosis, as well as the recording equipment and conditions. This rich dataset enables researchers to train and test machine learning models aimed at improving the accuracy and reliability of lung sound analysis, potentially aiding in early diagnosis and monitoring of respiratory diseases. The ICBHI Lung Sound Dataset stands out for its diversity and detailed annotations, making it a valuable resource for advancing respiratory sound analysis and improving clinical outcomes.

4.2 Chest Wall Database

The Mendeley Chest Wall Lung Sound Dataset[4] is an extensive compilation of respiratory sound recordings, curated to advance research in lung sound analysis. This dataset, accessi-

ble via the Mendeley Data repository, includes numerous recordings from subjects captured using high-quality digital stethoscopes placed on the chest wall. These recordings cover a variety of respiratory sounds, including normal breaths, wheezes, crackles, and other abnormal sounds. The dataset is meticulously annotated, providing details about the type of lung sound, the specific location on the chest where the recording was made, and information about the subjects such as age, gender, and health status (healthy or suffering from respiratory conditions). These comprehensive annotations are crucial for training and testing machine learning algorithms aimed at the automatic classification of lung sounds. The Mendeley Chest Wall Lung Sound Dataset is an invaluable resource for researchers focused on enhancing the accuracy and reliability of respiratory sound analysis. By offering a diverse range of recordings and detailed metadata, it supports the development of advanced diagnostic tools and algorithms, contributing significantly to the field of respiratory healthcare and technology.

4.3 SPR Sound Dataset

The SPRSound Dataset[5] is a comprehensive repository of respiratory sound recordings, designed to facilitate research in lung sound analysis and the development of diagnostic tools. The dataset comprises over 5,000 audio samples collected from 100 subjects, including both healthy individuals and patients with respiratory conditions such as asthma, bronchitis, and pneumonia. Recordings were made using high-quality digital stethoscopes, ensuring clear and accurate capture of lung sounds from six different locations on the chest wall. Each recording in the SPRSound Dataset is meticulously annotated, identifying different types of lung sounds such as normal breath sounds, wheezes, crackles, and other adventitious noises. The dataset includes extensive metadata, providing details such as the exact recording location on the chest, the subject's health status, age, gender, and smoking history. This comprehensive annotation is critical for training and evaluating machine learning algorithms for automatic lung sound classification and diagnosis. The dataset includes a variety of recordings that reflect different respiratory cycles and conditions, supporting the development of robust algorithms for detecting and analyzing respiratory sounds. The SPRSound Dataset is a valuable resource for researchers and clinicians aiming to improve diagnostic accuracy in respiratory health. By offering a rich collection of recordings and detailed annotations, it contributes significantly to the advancement of respiratory medicine and the development of better diagnostic tools and patient care practices.

4.4 HF Lung Dataset

The HF Lung Dataset is a specialized collection of respiratory sound recordings intended for advancing research in lung sound analysis and diagnostic technology. This dataset consists of high-fidelity audio samples obtained from individuals with various respiratory conditions, including asthma, chronic obstructive pulmonary disease (COPD), and pneumonia. Recorded using sensitive digital stethoscopes, the dataset ensures precise capturing of lung sounds from multiple positions on the chest wall. Comprising over 3,000 recordings from 200 subjects, the HF Lung Dataset provides a diverse range of respiratory sounds, annotated meticulously to distinguish between normal breath sounds, wheezes, crackles, and other abnormal sounds. Detailed metadata accompanies each recording, specifying factors such as the patient's age, gender, smoking history, and clinical diagnosis. This rich annotation facilitates the development and evaluation of machine learning algorithms tailored for automatic classification and diagnosis of respiratory conditions. Researchers and clinicians benefit from the dataset's comprehensive nature, enabling robust analysis of respiratory sound patterns across different diseases and patient demographics. The HF Lung Dataset plays a crucial role in enhancing the accuracy of respiratory health assessments, contributing to advancements in medical diagnostics and patient care strategies.

Chapter 5

Data Pre-Processing

This section explains the techniques to prepare the audio signals for feeding into our model. All the techniques have been itemized below:

5.1 Resampling

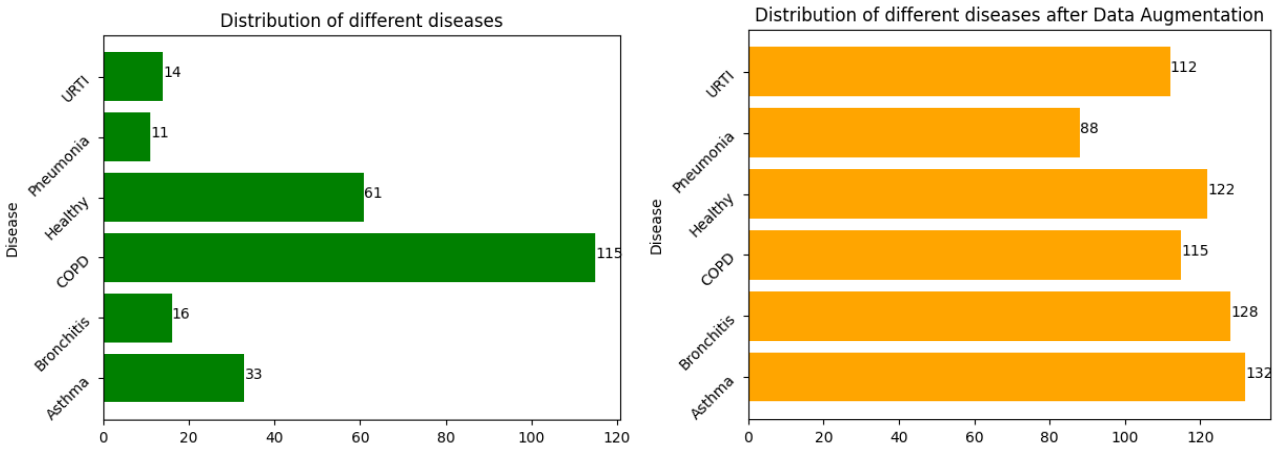
The audio signals from the ICBHI 2017 challenge database have a sampling rate in the range of [4, 44.1] KHz, while the signals from the other four databases have a sampling rate of 4 KHz. Hence, to bring uniformity to the signal properties, all the signals in the ICBHI database have been resampled to 4 KHz. It also becomes important because the highest frequency we want to deal with is 2 KHz(Wheezes), so a 4 KHz sampling rate is perfect.

5.2 Data Augmentation

After combining the three databases, we observe that audio signals from patients suffering from COPD are the majority class, while signals representing diseases like Asthma, Bronchitis, and Pneumonia are underrepresented. So, we superimpose ambient hospital noises on top of the original signals so that the resultant signal is closer to a real-life signal. We use noise signals of SNR in the range of [2, 5] dB and then superimpose it on the original signal. The data distribution representing different diseases before and after data augmentation can be found in Figure 5.1a and Figure 5.1b, respectively.

5.3 Amplitude Normalization

Amplitude normalization ensures that the signal amplitude is consistent across different samples or channels. This consistency is essential for accurate comparison, analysis, and interpretation of signals. We divide each amplitude value by the absolute of the highest amplitude value to



(a) Data distribution of diseases before data augmentation

(b) Data distribution of diseases after data augmentation

normalize the signals. If $\mathcal{S}_n$ represents a signal of length 10 seconds, the normalized signal $\tilde{\mathcal{S}_n}$ is given by

$$\tilde{\mathcal{S}_n} = \frac{\mathcal{S}_n - \mu_{\mathcal{S}_n}}{\sigma_{\mathcal{S}_n}}$$

5.4 Segmentation

The lung sound signals are now segmented into chunks of 10 seconds long to roughly cover the activity of inspiration and expiration cycle. While doing segmentation, instead of taking completely disjoint signals, we have an overlap of 50% between consecutive signals so that no information is lost on the edges.

Chapter 6

Feature Extraction Techniques

This section describes the various feature extraction techniques based on audio signals. Feature extraction acts as a bridge between raw audio data and its underlying meaning. By analyzing the audio signal for specific characteristics like pitch, rhythm, and spectral components, feature extraction helps us move beyond just the raw waveform. Two feature extraction techniques have been discussed below:

6.1 Mel-Spectrogram

Mel spectrograms are a powerful tool for visualizing and analyzing audio signals. They represent the short-time Fourier transform (STFT) of an audio signal on a Mel scale, which closely aligns with human auditory perception.

6.1.1 Short-Time Fourier Transform

The STFT decomposes a continuous audio signal into its constituent frequencies over time. It achieves this by applying a window function (e.g., Hann window) to short segments of the audio signal and then performing a Discrete Fourier Transform (DFT) on each windowed segment. The STFT component is calculated using equation 6.1.

$$X(\tau, f) = \sum_{n=0}^{N-1} x(n)w(n-\tau)e^{-2\pi jfn/N} \tag{6.1}$$

where

- X(τ, f): Represents the STFT coefficient at time index τ and frequency index f.
- $x(n)$: Discrete-time audio signal.

- $w(n)$: Window function applied to the audio segment (e.g., Hann window).
- τ: Time index (window position). It typically ranges from 0 to the number of windows - 1.
- f: Frequency index. It typically ranges from 0 to N/2 for real-valued signals due to the Nyquist-Shannon sampling theorem.
- N: Window size (number of samples in the window).
- j: Imaginary unit.
- $e^{-2\pi jfn/N}$: Complex exponential term for frequency shifting.

6.1.2 Mel Scale

The Mel scale is a non-linear transformation that approximates human auditory perception. High-frequency sounds are perceived closer together than low-frequency sounds. The Mel scale compresses high frequencies while expanding low frequencies, providing a more perceptually relevant representation. The conversion between frequency in Hz and frequency in mels is related by equation 6.2

$$mel(f) = 2595 \times log10\dot{(}1 + \frac{f}{700}) \tag{6.2}$$

where

- f: Frequency in Hz
- mel: Frequency in Mels

6.1.3 Mel Filter Banks

Mel filter banks are a set of band-pass filters spaced according to the Mel scale. These filters are applied to the magnitude spectrum (absolute value) of the STFT to capture the energy within each Mel band.

6.1.4 Mel Spectrogram

The Mel spectrogram is a visual representation where the x-axis represents time (corresponding to the time indices of the STFT), the y-axis represents frequency on the Mel scale, and the colour intensity at each (time, frequency) coordinate indicates the energy (loudness) within that Mel band at that particular time instance.

6.2 Mel-frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCCs) are a powerful feature extraction technique widely used in speech and audio processing tasks. They capture the characteristics of an audio signal based on human auditory perception. The steps involved in calculating MFCCs are as follows:

6.2.1 Preprocessing

- Framing: The audio signal is segmented into short, overlapping frames (typically 20-40 milliseconds).
- Windowing: A window function (e.g., Hann window) is applied to each frame to reduce spectral leakage (artefacts caused by abrupt signal truncation).

6.2.2 Mel Scale Conversion

After windowing, we apply the mel scale conversion and mel filter banks described in sections 6.1.2 and 6.1.3, respectively.

6.2.3 Discrete Cosine Transform

The DCT is applied to the logarithm of the Mel filter bank outputs. This emphasizes the cepstral coefficients that correspond to the slowly varying spectral envelope of the signal, which is important for speech recognition. The mathematical equation for DCT is given in equation 6.3.

$$C(n) = \sum_{k=0}^{N-1} \log(X_m(k)) \cos\left(\pi n(k + 0.5)/N\right) \tag{6.3}$$

where

- $C(n)$: nth cepstral coefficient
- $X_m(k)$: Mel filter bank output for k_{th} frequency bin in m_{th} frame
- N: Number of DCT coefficients

Chapter 7

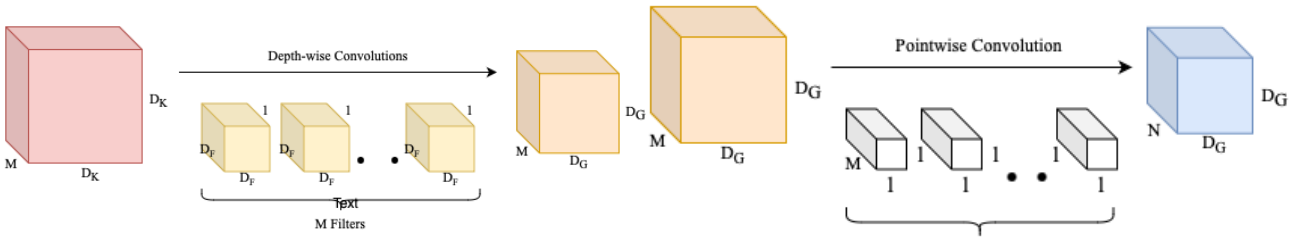
Deep Learning Model

Deep learning models, particularly neural networks with multiple layers, are employed in research and applications due to their unparalleled ability to automatically learn intricate patterns and representations from vast and complex data. Unlike traditional machine learning algorithms that often require handcrafted features, deep learning models can discover hierarchical representations from raw data, enhancing performance in various tasks across multiple domains.

This section explores the lightweight deep learning architecture and compares it to a network with regular CNN architectures. Then we move on to the Wav2Vec2 Architecture by Facebook, which is used for Automatic Speech Recognition, and how we change the architecture to use it for our use case. These are the two model architectures that have been explored in this project work.

7.1 Depthwise Separable Convolution

Regular convolutions and depthwise separable convolutions perform feature extraction on images, but the latter achieves this with significantly fewer parameters. Regular convolution learns a single set of weights that combines spatial and channel-wise information. This can be



(a) Depthwise convolutions (b) Pointwise convolutions

redundant for tasks where the spatial and channel-wise information can be separated. Depthwise convolution independently captures spatial information for each channel, leading to a more efficient use of parameters. The pointwise convolution combines these features across channels with a much smaller set of filters than a regular convolution.

Let us consider a volume of $D_h \times D_w \times C$ where D_h is the height of the volume, D_w is the width of the volume, and C is the number of channels. In the case of a regular convolution, a kernel with $D_k \times D_k$ size has ${D_k}^2 \times C$ parameters. For F filters, the number of parameters becomes ${D_k}^2 \times C \times F$. On the other hand, in depthwise separable convolutions, we have $D_k \times D_k \times C$ depthwise features and $1 \times 1 \times C \times F$ pointwise features. To do a comparison, we see that,

$$\begin{aligned}\frac{CNN}{DSC} &= \frac{D_k^2 \times C \times F}{D_k^2 \times C + C \times F} \\ &= \frac{D_k^2 \times C \times F}{D_k^2 \times C + C \times F} \\ &= \frac{D_k^2 \times F}{D_k^2 + F}\end{aligned}$$

Since D_k and F represent positive quantities, this ratio is always greater than 1 for $D_k > 1$ and $F > 1$. Hence, the number of parameters in a regular convolution network is higher than in depthwise separable convolution.

7.2 Transformer Architecture

Wav2Vec 2.0[2] utilizes a transformer-based encoder architecture. The encoder consists of stacked self-attention layers with residual connections and layer normalization. Positional encodings are added to the input to capture the temporal order of audio samples. The model is trained using a contrastive learning objective called masked language modelling (MLM) for audio. As masked language modelling works for text, the model predicts masked audio tokens based on the surrounding context. This objective encourages the model to learn general audio representations that capture the relationships between different parts of an utterance. Wav2Vec 2.0 significantly improves over its predecessor by utilizing a transformer-based architecture, which is more powerful for capturing long-range audio dependencies than the convolutional architecture used in Wav2Vec 1.0. It has also introduced the masked language modelling objective, which allows unsupervised pre-training on large amounts of unlabeled audio data and has employed layer normalization for improved stability during training. The model architecture is

represented in Figure 7.2.

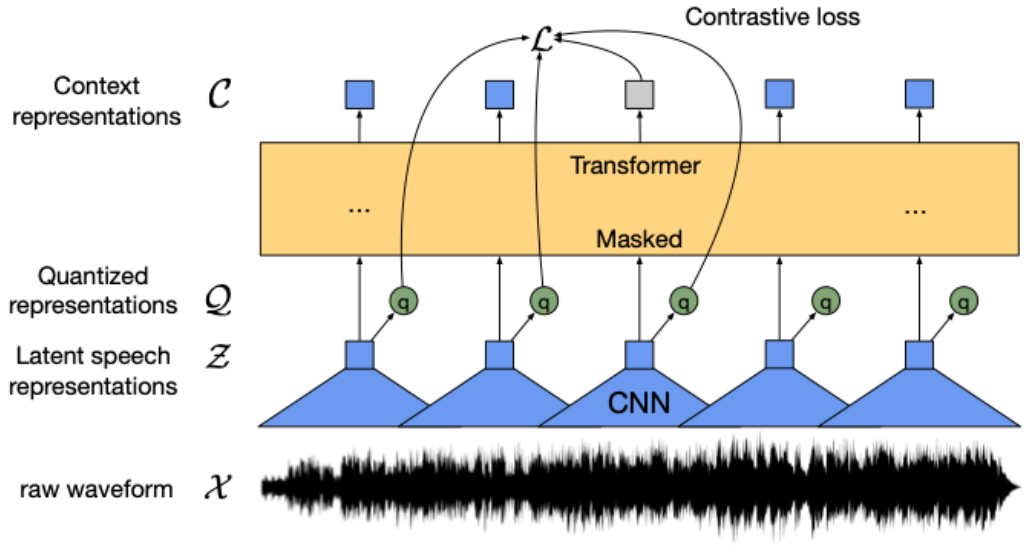


Figure 7.2: Wav2Vec 2.0 Architecture [2]

7.2.1 Transformer Pruning

Wav2Vec2 architecture comprises a feature extractor part and a feature encoder part. Feature Extractor part has 7 one dimensional convolutional layers, each comprising 512 filters, so the resulting window size becomes 20msec. The feature encoder has 12 encoder blocks. The total number of parameters can be found in Table 7.1

Module	**No. of Parameters**
Feature Extraction Layer	4.2M
Feature Encoder Layer	89.7M
Total Parameters	**94.3M**

Table 7.1: Number of parameters in different modules of Wav2Vec2 architecture

We can see that the encoder blocks account for over 95% of parameters. One simple technique to decrease the number of parameters is to decrease the number of encoder blocks and study how it affects the model performance. In this work, we have decreased the number of encoder blocks from 12 to 8.

One major contributor of parameters is the use of feed forward dense neural network. In the original Wav2Vec2 architecture, these dense layers contribute to over 63% parameters in

feature encoder layers and over 60% in all of Wav2Vec2 parameters. Hence, we must try to decrease the number of parameters in these dense layers.

7.2.2 AutoEncoder Feed Forward Neural Network

We propose an autoencoder architecture to replace the feed-forward dense layers. This translates to introducing a bottleneck layer between a layer and its projection layer. In the context of the Wav2Vec2 layer, the hidden side of embeddings is 768, which is then projected to 3072 dimensions and then back to 768. We place a bottleneck layer in between these layers to decrease the number of parameters.

Theorem 7.1 *Let layers A and B have x and y number of neurons. Then the number of neurons in a bottleneck layer, which guarantees a lower number of parameters than not using a bottleneck layer, is given by $\lfloor \frac{x \cdot y}{x+y} \rfloor$*

Proof: If we have two layers with x and y number of neurons. Then, the number of parameters required to fully connect these two layers is $x \cdot y$. Let the number of neurons in the bottleneck layer be c. Then, the number of parameters in the new architecture becomes $x \cdot c + c \cdot y$.

$$\begin{aligned} &\implies x \cdot c + c \cdot y \leq x \cdot y \\ &\implies c \cdot (x + y) \leq x \cdot y \\ &\implies c \leq \frac{x \cdot y}{x + y} \end{aligned}$$

Since the number of neurons cannot be fractional, hence we use the greatest integer function value i.e. $\lfloor \frac{x \cdot y}{x+y} \rfloor$ □

In our Wav2Vec2 architecture, the feed-forward neural network has three layers with sizes 768, 3072, and 768. By using the result in theorem 7.1, we find that the maximum number of neurons that we can put in our bottleneck layer is 614.

No. of neurons	**% decrease in parameters**
64	65.9%
128	61.7%
256	53.4%
512	36.7%

Table 7.2: Number of neurons in bottleneck layer and corresponding decrease in number of parameters

Chapter 8

Experimental Setup

We have created our experimental setup on two tasks. These are listed below:

1. **Task 1**: The Adventitious Sound Detection task is a 4-class classification problem, where the four classes are Normal, Crackle, Wheeze, and Both Crackle and Wheeze.

2. **Task 2**: The Disease Detection Task is a 4-class classification problem as well where the four classes are Asthma, COPD(Chronic Obstructive Pulmonary Disorder), Healthy, and Other, where other signifies the presence of Bronchitis, Pneumonia, URTI(Upper Respiratory Tract Infection), and Lung Fibrosis.

There are two types of experiments done on these two tasks. First is the performance of a lightweight CNN architecture and how it compares with a regular CNN architecture with the same number of layers. Second is the performance of transformer based architectures, like Facebook's Wav2Vec2 Base model, and our Wav2Vec2 model with custom configuration and trained exclusively on heart and lung sounds only. We compare the performances across different models and observe if there is a trade-off between model parameters and model accuracy.

8.1 Lightweight CNN architecture

For our lightweight CNN architecture, we use a combination of regular CNN layers and depth-wise separable convolution layers. The model architecture is shown in Figure 8.1

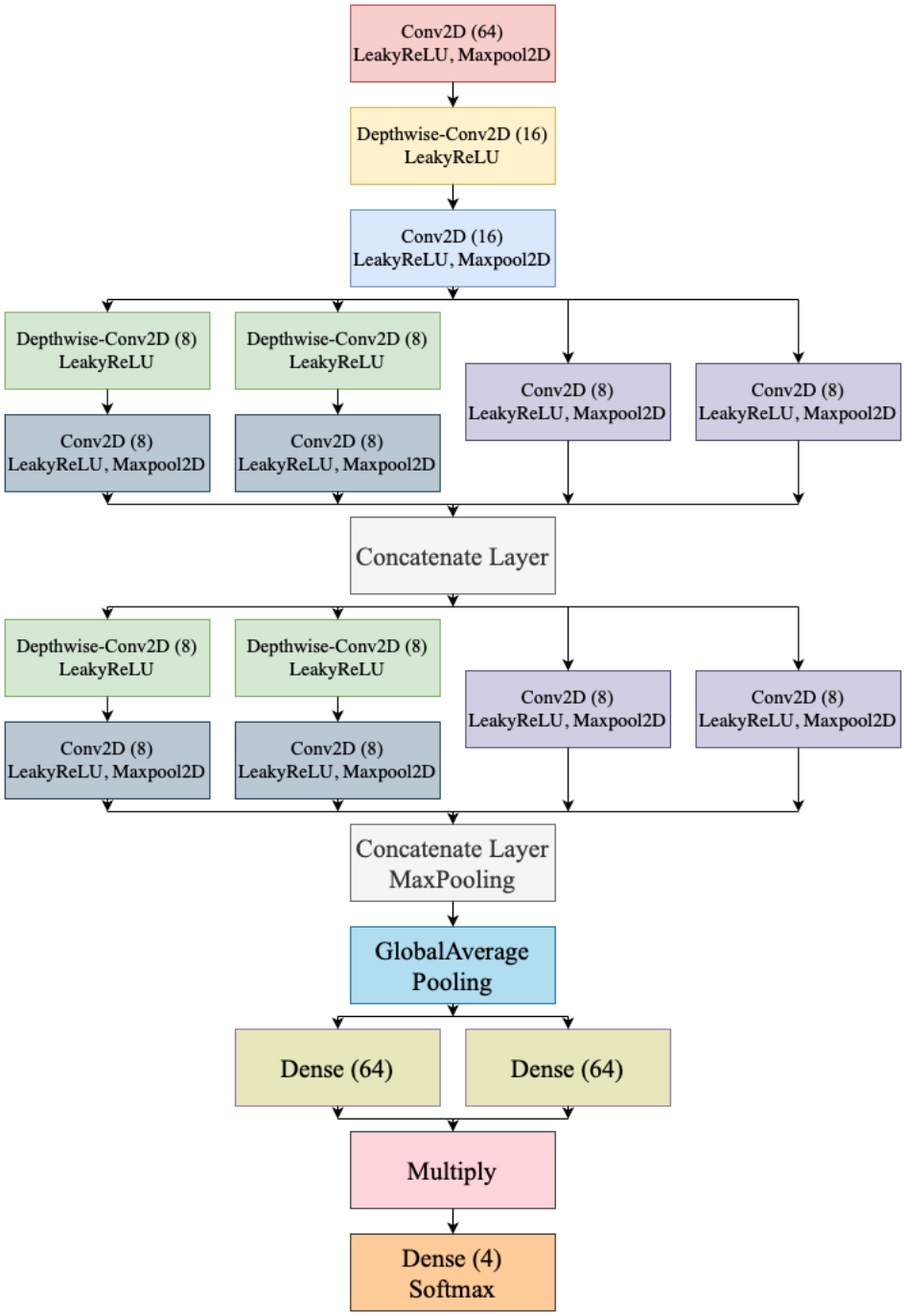


Figure 8.1: Lightweight CNN model architecture

Multiple feature extraction techniques have been used. We record observations for raw signal, mel-spectrogram, and MFCC coefficients. We have considered the window size of 8 seconds. The choice of 8 seconds is made as it can roughly cover at least one respiratory cycle. The model training experiments were run on 8 Tesla V100-SXM2-32GB GPU cores provided

by the Department of Computer Science and Automation, Indian Institute of Science, at DGX server. Our model remains the same for both tasks. The database administrators gave the train and test split for ICBHI data, and the same split has been used here. A train-test split was created for the remaining two databases before any experimentation, and that same split was used for all the experiments. This is done to ensure a fair comparison between different experiment hyperparameters. Further, the test files are randomly distributed among 10 sub-test folders. Experiments that involve mel-spectrogram and MFCC coefficients involve various hyper-parameters. These are reported in Table 8.1.

Sampling Rate	4 KHz
Length of the FFT window	1024
Number of samples between successive frames	128, 256, 512
Frame Length	1024
Window Type	hann
Number of Mel bands	64
Number of MFCCs returned	64
Discrete cosine transform (DCT) type	2

Table 8.1: Hyperparameters for experimental setup

8.2 Pretraining of Wav2Vec2 Architecture

The original Wav2Vec2 model was trained on speech sounds. To better match our use case, we have trained the original Wav2Vec2 model majority on lung sounds, and some heart sounds. Including heart sounds is done because, during lung sound examination, some heart sounds are superimposed on the lung sounds.

Wav2Vec 2.0 leverages several key components to learn powerful speech representations. These are listed below:

1. **Feature Extractor:** This module processes the raw audio waveform and extracts meaningful features. In the original wav2vec 2.0 architecture, we have 7-layer stacks of convolutional layers. Since we have resampled our audio files to 4KHz, we need only a 5-layer convolutional layer network to achieve the 50Hz resolution or 20ms window. Combined with appropriate filters, these layers capture local patterns in the audio signal. These layers can learn to detect fundamental audio components like phonemes.

2. **Context Network:** This network captures long-range dependencies within the extracted features. The core component here is a stack of Transformer encoder layers. This layer consists of two sub-layers. The first is a Multi-Head Self-Attention, which allows the model to attend to different aspects of the input features simultaneously, capturing relationships between features at different positions. Second is a Feed-Forward Network, which adds non-linearity to the model, allowing it to learn complex relationships between features. This is the layer where we introduce a bottleneck layer and try to reduce the overall number of parameters in our model.

3. **Contrastive Loss Function:** This drives the model's learning process. It encourages the model to learn representations that differentiate between positive and negative pairs of audio segments:

   - Positive Pairs: These represent audio segments with similar content (e.g., consecutive segments from the same audio).
   - Negative Pairs: These represent audio segments with dissimilar content (e.g., segments from audios of different classes).

   The model learns to encode similar audio segments into similar latent representations in the embedding space while pushing dissimilar segments apart. The formula for the same can be found in equation 8.1.

$$L(z_i, z_j) = -\log \frac{\exp\left(\text{sim}(z_i, z_j)/\tau\right)}{\sum_{k \neq i} \exp\left(\text{sim}(z_i, z_k)/\tau\right)} \tag{8.1}$$

   where,

   - $L(z_i, z_j)$ is the loss value for the pair of latent representations z_i and z_j,
   - τ is a hyperparameter that controls the influence of the disparity between positive and negative pairs. Higher temperatures lead to a softer distinction,
   - $sim(z_i, z_j)$ is the similarity score between the two latent representations z_i and z_j. This can be calculated using various similarity functions, such as the cosine similarity.

The different hyperparameters used to pretrain our wav2vec 2.0 architecture are given in Table 8.2.

Hyperparameter	Value
Sampling Rate	4000Hz
Audio chunk length	10sec
Learning Rate	0.0001
Number of distractors	100
τ	0.999995
No of encoder blocks	8
No of neurons in bottleneck layer	64, 128, 256, 512

Table 8.2: Hyperparameters used in Wav2Vec2 Pretraining

8.3 Finetuning of Wav2Vec2 Architecture

After pretraining Wav2Vec2 architecture, we finetune this architecture on the ICBHI Dataset for Adventitious Sound Detection, and ICBHI Dataset and ChestWall Database for Disease Detection. We only add a classification head after the wav2vec2 architecture. We have four neurons in the output layer for both tasks.

Chapter 9

Results

This section summarises the results observed in the four experiments. These four experiments are:

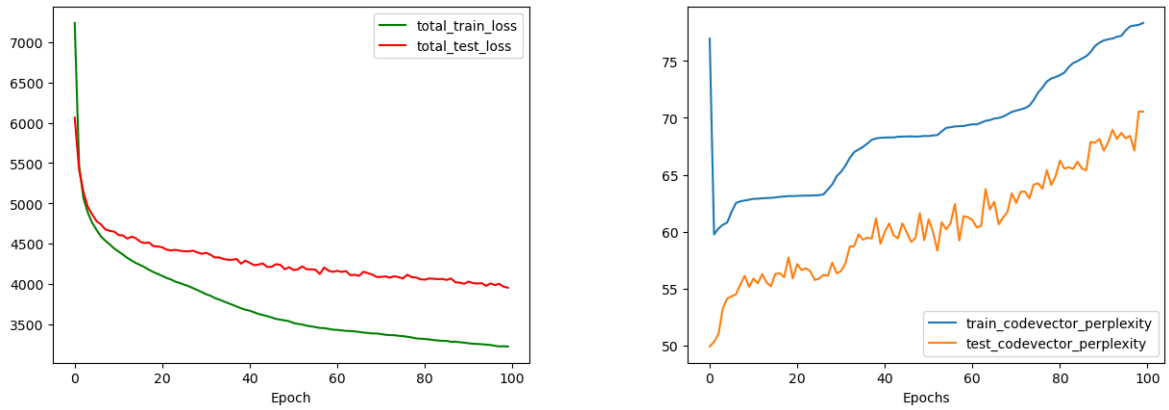
- Task 1 using lightweight CNN architecture
- Task 2 using transformer architecture
- Task 1 using lightweight CNN architecture
- Task 2 using transformer architecture

In lightweight CNN architecture, we compare the results of using different hyperparameters like the number of mel filter banks and the number of samples between successive frames. We also compare these numbers with a similar regular CNN architecture. In the transformer architecture, we compare the results with Facebook's `wav2vec2-base`. We also compare the results between pretrained models of different bottleneck layers on our two downstream tasks.

9.1 Pretraining Wav2Vec2

We discuss the results of five pretraining experiments on our wav2vec2 architecture. We trained five Wav2Vec2 architectures. The database consisted of 10-second samples from the SPR Sound and HF Lung V1 databases. These 5 architectures are:

- Wav2Vec2 architecture (wav2vec2-pretrained)
- Wav2Vec2 architecture with 64 neurons in bottleneck layer (wav2vec2-pretrained-64)
- Wav2Vec2 architecture with 128 neurons in bottleneck layer (wav2vec2-pretrained-128)



(a) Train and test loss for wav2vec2-pretrained-64 model pretraining

(b) Code Vector Perplexity for wav2vec2-pretrained-64 model pretraining

Figure 9.1: Training Loss Plot and Perplexity PLot

- Wav2Vec2 architecture with 256 neurons in bottleneck layer (wav2vec2-pretrained-256)
- Wav2Vec2 architecture with 512 neurons in bottleneck layer (wav2vec2-pretrained-512)

The loss values of test samples while training these architectures are given in Table 9.1.

No. of neurons in bottleneck layer	**Loss Value at Epoch 0**	**Loss Value at Epoch 200**
0	5943.44	2924.09
64	11489.24	6129.54
128	9328.43	5723.35
256	8321.87	4749.31
512	7625.21	4298.71

Table 9.1: Loss Values for different Wav2Vec2 architectures

As is evident from Table 9.1, there is a decrease in the loss value between Epoch 0 and Epoch 200, but the loss value is still not low enough to call the model converged. For the model 'wav2vec2-pretrained-64', the training loss plot and perplexity are depicted in figures 9.1a and 9.1b, respectively.

9.2 Adventitious Sound Detection

In this section, we summarise the results of our experiments on crackle and wheeze detection in recorded lung sounds. This experiment aims to give lung sound as an input and predict one of the 4 classes. These classes are healthy lung sounds, crackles, wheezes, and crackles and wheezes. Section 9.2.1 refers to the results of the lightweight CNN architecture on this task, and section 9.2.2 refers to the results of wav2vec2 models on this task.

9.2.1 Lightweight CNN architecture

Table 9.2 summarises the accuracy values on lightweight CNN architecture and regular CNN architecture on different feature extraction techniques for different hop lengths on adventitious sound detection tasks.

Feature Extraction Technique	**Hop Length**	**Accuracy of light CNN**	**Accuracy of regular CNN**
Raw Signal	-	67.1%	67.3%
Mel-Spectrogram	128	**78.7%**	77.9%
	256	77.5%	78.2%
	512	77.4%	77.1%
MFCC coefficients	128	77.8%	78.1%
	256	76.6%	77.2%
	512	77.1%	77.4%

Table 9.2: Accuracy of different feature extraction techniques on light CNN and regular CNN for sound detection

9.2.2 Wav2Vec2 Architecture

Table 9.3 summarises the accuracy values on different wav2vec2 architectures for adventitious sound detection tasks.

Wav2Vec2 Architecture	Accuracy
`wav2vec2-base`	29.4%
wav2vec2-pretrained	36.4%
wav2vec2-pretrained-64	38.8%
wav2vec2-pretrained-128	36.1%
wav2vec2-pretrained-256	38.4%
wav2vec2-pretrained-512	**38.9%**

Table 9.3: Accuracy of different feature extraction techniques on various wav2vec2 architectures for sound detection

9.3 Disease Detection

In this section, we summarise the results of our experiments on asthma and COPD detection in recorded lung sounds. This experiment aims to give lung sound as an input and predict one of the 4 classes. These classes are healthy, asthma, COPD, and other diseases (Bronchitis, Pneumonia, URTI, and Lung Fibrosis). Section 9.3.1 refers to the results of the lightweight CNN architecture on this task, and section 9.3.2 refers to the results of wav2vec2 models on this task.

9.3.1 Lightweight CNN architecture

Table 9.4 summarises the accuracy values on lightweight CNN architecture and regular CNN architecture on different feature extraction techniques for different hop lengths for disease detection.

Feature Extraction Technique	Hop Length	Accuracy of light CNN	Accuracy of regular CNN
Raw Signal	-	73.8%	70.7%
Mel-Spectrogram	128	85.1%	83.5%
	256	**85.9%**	84.9%
	512	84.4%	82.7%
MFCC coefficients	128	84.9%	84.7%
	256	83.9%	85.4%
	512	82.3%	81.4%

Table 9.4: Accuracy of different feature extraction techniques on light CNN and regular CNN for disease detection

9.3.2 Wav2Vec2 Architecture

Table 9.5 summarises the accuracy values on different wav2vec2 architectures for disease detection tasks.

Wav2Vec2 Architecture	**Accuracy**
`wav2vec2-base`	77.2%
wav2vec2-pretrained	78.4%
wav2vec2-pretrained-64	79.1%
wav2vec2-pretrained-128	80.8%
wav2vec2-pretrained-256	**81.6%**
wav2vec2-pretrained-512	80.6%

Table 9.5: Accuracy of different wav2vec2 architectures on light CNN and regular CNN for disease detection

Chapter 10

Observations

In this section, we make observations based on the results we got in Section 9.

10.1 Pretraining Wav2Vec2

Table 9.1 shows the test data loss values. We can notice that although there is a significant decrease, the model is still far from convergence. Training Wav2Vec 2.0 involves processing massive amounts of audio data, and a single GPU hinders the model's exploration of the vast parameter space. This exploration is crucial for finding the optimal configuration that minimizes loss. A single GPU also restricts the batch size, potentially leading to unstable gradients during training. These unstable gradients can make it harder for the model to learn the complex relationships within the audio data, impacting the capture of diversity within the representations. With only a single GPU, parallelization efficiency might be limited. This can lead to underutilized computational resources and hinder the model's ability to fully leverage the available processing power to optimize perplexity, a measure of how well the model predicts the next element in the audio sequence.

10.2 Adventitious Sound Detection

Table 9.2 on Page 30 shows the results of using a lightweight CNN architecture for this task. It shows that the raw signal sound achieves an accuracy of 67.1% on the lightweight CNN and 66.29% on regular CNN. The value for the raw signal is the lowest among all the feature extraction techniques. It is also important to state that at 4 KHz sampling rate and 10 sec window length, there are 40000 samples which is a lot to process. Hence, the training time for the raw signal was the highest at 4 hours for 100 epochs and 128 batch sizes.

The performance of mel-spectrogram and MFCC regarding model accuracy and training

time is very similar. There is a sharp increase in accuracy values from the raw signal as we provide our model with more meaningful features. Also, there is a drastic decrease in features, from 40000 to 10000. This is responsible for a shorter training time than raw signal as the model was trained in 30 minutes.

Coming to wav2vec2 architecture, we notice a sharp decline in the accuracy values, as evident in 9.3. The best performance is given by wav2vec2-pretrained-512 architecture with an accuracy value of 38.9%. This can be attributed to the explanation given in Section 10.1. Since the model has not converged yet, the diversity in data is not learnt to the full extent, and hence the low accuracy values.

Please note that lightweight CNN architecture combined with the mel-spectrogram feature extraction technique with a hop of 128 samples between consecutive frames leads to the best performance among the lot.

10.3 Disease Detection

Table 9.4 on Page 32 shows that the raw signal sound achieves an accuracy of 73.8% on lightweight architecture and 70.7% on regular CNN architecture. The values for the raw signal technique are the lowest among all the feature extraction techniques, just as was the case in the previous section. The training time for the raw signal was the highest at 4.5 hours for 100 epochs and 128 batch size.

It is also observed that the performance and training time of the mel-spectrogram and MFCC feature extraction methods are almost comparable. The best accuracy values for mel-spectrogram and MFCC were 85.9% and 84.9%, respectively.

There is a noticeable improvement in the accuracy values in the wav2vec2 architectures when it comes to the disease classification tasks compared to the adventitious sound classification tasks. This might be because the pre-training focuses on capturing general characteristics of lung sounds, potentially including healthy and abnormal variations. While this is beneficial for identifying broad disease categories, it might not provide enough detail for pinpointing specific adventitious sounds like crackles or wheezes, which have more subtle acoustic signatures. For improved performance on these finer-grained classifications, the model might benefit from additional training specifically focused on the acoustic properties of these adventitious sounds.

Chapter 11

Conclusion and Future Work

In this work, we tried to check different feature extraction techniques and how they perform on our deep learning models. These models ranged from simple CNN architectures to complex Wav2Vec2 architectures. The former, combined with an efficient parameter reduction technique, resulted in better performance than regular CNNs on both tasks in various experimental settings. We also notice that mel-spectrogram outperforms all other feature extraction techniques in both tasks. The numbers for wav2vec2 architectures in Task 1 did not look encouraging, but those can be partially attributed to the limited availability of data and the detailed intricacies involved in identifying the acoustic properties of crackles and wheezes.

The future work includes incorporating the sound detection and disease detection task hierarchically, as many literary sources indicate a direct correlation between the type of adventitious sound and lung disease caused by its presence. The pretraining of our wav2vec2 architecture can also be taken up as future work to bring the model loss to convergence and then evaluate the model on downstream tasks. This can include changes in techniques adopted to pretrain our model or collect more data as it will help the model see more diverse sound types.

Appendix A

Lightweight CNN Experiments

A.1 Asthma v/s Healthy Disease Classification

This section summarises the results of the experiments done for a two class classification where the two classes are Asthma and Healthy.

A.1.1 Data Oversampling

Feature Extraction Technique	**Hop Length**	**Accuracy**	**Senstivity**
Raw Signal	-	72.12%	64.1%
Mel-Spectrogram	128	80.97%	57.69%
	256	80.53%	58.97%
	512	80.97%	66.67%
MFCC coefficients	128	79.65%	61.54%
	256	**84.07%**	**74.36%**
	512	80.09%	64.1%

Table A.1: Accuracy and Recall of different feature extraction techniques on light CNN for 2 class disease detection when data is oversampled

A.1.2 Data Undersampling

Feature Extraction Technique	**Hop Length**	**Accuracy**	**Senstivity**
Raw Signal	-	73.89%	55.13%
Mel-Spectrogram	128	80.97%	64.1%
	256	85.4%	69.23%
	512	**88.5%**	**73.08%**
MFCC coefficients	128	80.53%	69.23%
	256	82.74%	67.95%
	512	81.86%	64.1%

Table A.2: Accuracy and Recall of different feature extraction techniques on light CNN for 2 class disease detection when data is undersampled

A.2 Asthma v/s COPD v/s Healthy Disease Classification

This section summarises the results of the experiments done for a three class classification where the three classes are Asthma, COPD and Healthy.

A.2.1 Data Oversampling

Feature Extraction Technique	**Hop Length**	**Accuracy**	**Senstivity**
Raw Signal	-	81.79%	76.17%
Mel-Spectrogram	128	85.14%	70.76%
	256	**85.89%**	73.67%
	512	84.39%	**75.79%**
MFCC coefficients	128	84.44%	71.06%
	256	80.09%	72.70%
	512	81.59%	70.68%

Table A.3: Accuracy and Recall of different feature extraction techniques on light CNN for 3 class disease detection when data is oversampled

A.2.2 Data Undersampling

Feature Extraction Technique	**Hop Length**	**Accuracy**	**Senstivity**
Raw Signal	-	79.99%	67.11%
Mel-Spectrogram	128	78.69%	70.89%
	256	80.69%	73.10%
	512	77.24%	72.98%
MFCC coefficients	128	77.59%	70.81%
	256	**80.94%**	**77.61%**
	512	75.74%	70.43%

Table A.4: Accuracy and Recall of different feature extraction techniques on light CNN for 3 class disease detection when data is undersampled

Bibliography

[1] GBD Chronic Respiratory Disease Collaborators. Prevalence and attributable health burden of chronic respiratory diseases, 1990-2017: a systematic analysis for the Global Burden of Disease Study, Lancet Respir Med., 2017. 1

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. vi, 20, 21

[3] Coimbra M. Faustino P, Oliveira J. Crackle and wheeze detection in lung sound signals using convolutional neural networks, 2021. 8

[4] Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief*, 35:106913, 2021. ISSN 2352-3409. doi: https://doi.org/10.1016/j.dib.2021.106913. URL https://www.sciencedirect.com/science/article/pii/S2352340921001979. 11

[5] Fu-Shun Hsu, Shang-Ran Huang, Chien-Wen Huang, Yuan-Ren Cheng, Chun-Chieh Chen, Jack Hsiao, Chung-Wei Chen, and Feipei Lai. An update on a progressively expanded database for automated lung sound analysis, 2021. 12

[6] Shubham Kulkarni, Hideaki Watanabe, and Fuminori Homma. Self-supervised audio encoder with contrastive pretraining for respiratory anomaly detection. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5, 2023. doi: 10.1109/ICASSPW59220.2023.10193030. 9

[7] Truc Nguyen and Franz Pernkopf. Lung sound classification using snapshot ensemble of convolutional neural networks. volume 2020, 04 2020. doi: 10.1109/EMBC44109.2020.9176076. 9

[8] Jemisha Patel and Milin Patel. Different transfer learning approaches for recognition of lung sounds: Review. pages 738–742, 02 2022. doi: 10.1109/ICAIS53314.2022.9742754. 8

[9] Rocha BM et al. An open access database for the evaluation of respiratory sound classification algorithms, 2019. 11