Diagnosis of respiratory sounds using Deep Neural Networks

Kuljeet Singh
(kuljeet.singh@wipro.com)
Solution Architect, Wipro Engineering Edge
East Brunswick, NJ 08816

Abstract

Respiratory diseases are on the rise around the world and the impact of COVID-19 has highlighted the need for early and better diagnosis of respiratory diseases. The growing air pollution from vehicles, wildfires, coal-burning power plants and other natural and non-natural causes, particularly in the developing world is leading to more deaths due to respiratory problems leaving many in need of diagnosis and treatment. The use of machine learning for preliminary analysis and diagnosis of diseases is evolving rapidly, particularly in the area of analysis of medical images to help sort through and analyze hundreds of X-ray, CT-Scan or MRI images to highlight the area affected by a potential medical condition and suggest a possible diagnosis. This has allowed timely detection of potentially life-threatening diseases, reduced diagnosis time, improved efficiency and better coverage. These medical aids are also being integrated into the medical scanning equipment and related software to further improve the diagnostic process.

This paper attempts to expand the use of deep learning as a method of assisting in the early diagnosis of respiratory diseases using audio recordings. The paper describes an approach for analyzing lung sounds captured using an electronic stethoscope from different parts of a patient's chest wall. The paper describes the processes used to extract features from the sound signals, dataset preparation, neural network architectures evaluated and the prediction results.

Introduction

Chronic respiratory diseases are one of the leading causes of morbidity and mortality worldwide. In many cases, such deaths are preventable with early diagnosis. Diseases like chronic obstructive pulmonary disease (COPD), asthma, bronchitis and pleural effusion can lead to a reduced quality of life, disability or even death. The increase in lung diseases in children under the age of 5, including infectious processes and chronic conditions such as asthma are among the most common causes of mortality affecting about 14% of children and rising. Among the most common causes of the rise in chronic respiratory conditions are degrading air quality due to pollution, viral infections, exposure to toxic work or living environments and primary or secondary tobacco smoke inhalation. Data from the WHO and Global Burden of Chronic Respiratory Diseases[1] (GBD) study show that nine out of 10 people are exposed to high levels of air pollutants with up to 3.2 million deaths due to COPD and 495,000 deaths due to asthma in a year.

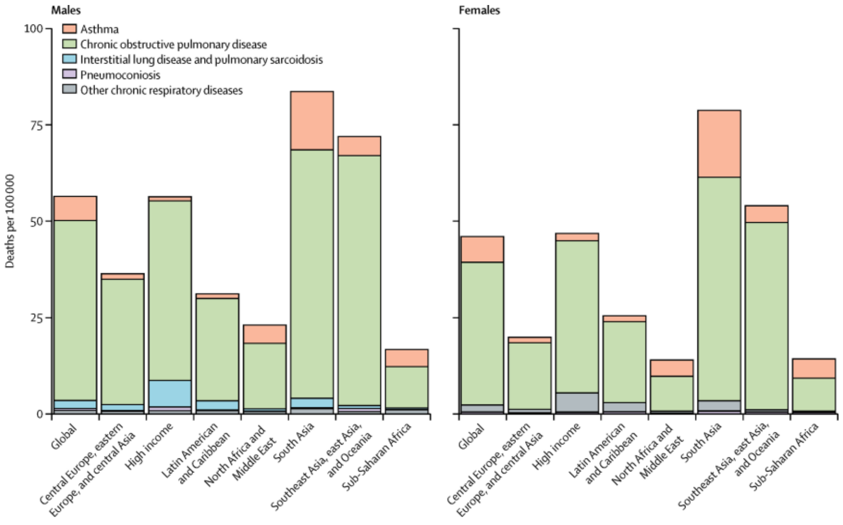


Figure 1: Region-wise spread of respiratory diseases[3]

The probability of contracting a respiratory ailment varies dramatically with the age and gender of the patient with asthma being a leading cause in children between the ages of 5 and 9, whereas COPD is the major contributor in adults and the elderly with males being more vulnerable than females.

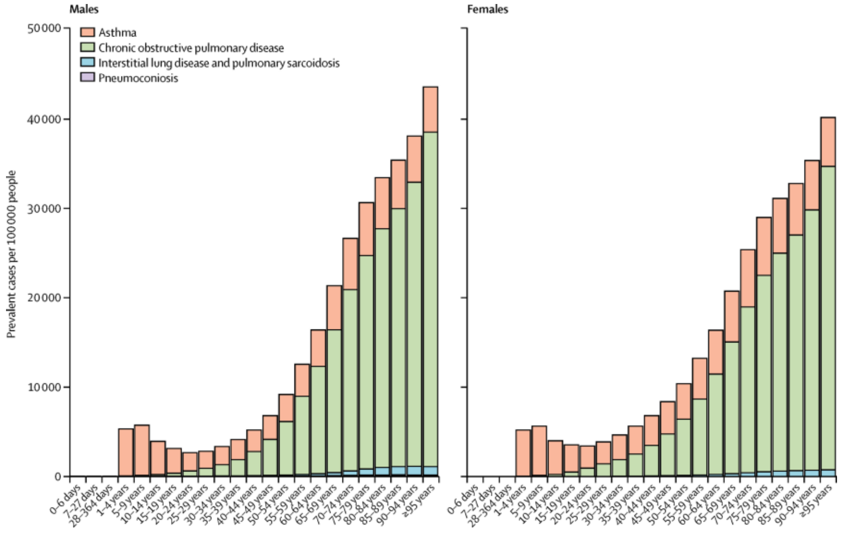


Figure 2: Age-wise spread of respiratory diseases[3]

Among the major factors contributing towards the rise of respiratory diseases, the most prevalent risk is smoking with ambient air pollution and occupational exposure being other leading factors. Demographically, people living in underdeveloped and developing countries with large population bases are especially vulnerable to respiratory diseases due to the lack of clean environment, availability and access to medical treatment and limited coverage of health care services.

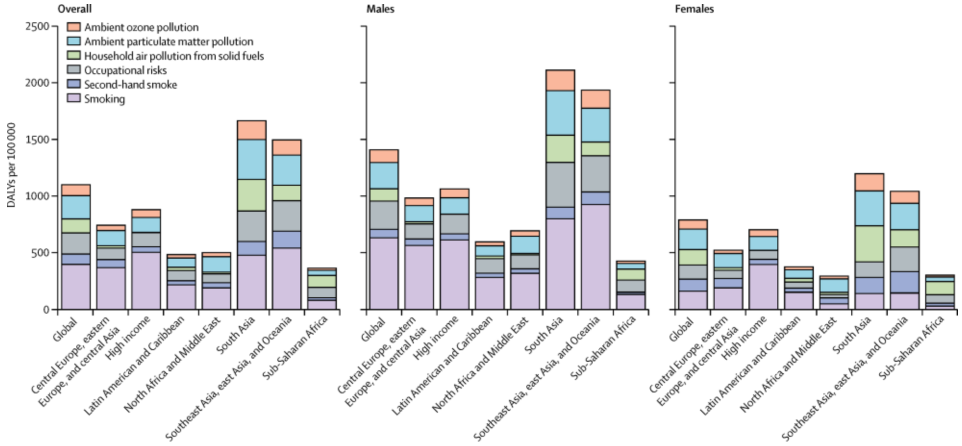


Figure 3: Prime contributors to respiratory conditions globally[3]

COVID-19 has further highlighted the need for better and quicker diagnosis of respiratory symptoms. The traditional methods used for respiratory diagnosis include pulmonary function tests, Check X-ray, CT scan and arterial blood gas analysis. The use of machine learning as a diagnosis aid is still in its infancy but is already showing promise. In this paper, I will discuss an approach for using deep learning techniques to analyze the audio of respiratory sounds for diagnosis of the seven most common medical conditions including asthma, pneumonia, bronchitis, pleural effusion, lung fibrosis, heart failure and chronic obstructive pulmonary disease (COPD).

Background and Approach

In this approach, we focus on analyzing respiratory sound recordings taken during patient examination by placing an electronic stethoscope at various points over the chest wall. The audio recordings are then processed using various filters to enhance their quality in various frequency ranges. Audio features are extracted from samples using various techniques like Mel-Spectrogram, MFCC and short-time Fourier transform(STFT). Once extracted, these features are converted to a pixel representation and stacked on top of each other to form a multi-channel image representing various aspects of the audio signal over a time frame. These representative images are then used to analyze and classify the recordings to one or more respiratory conditions.

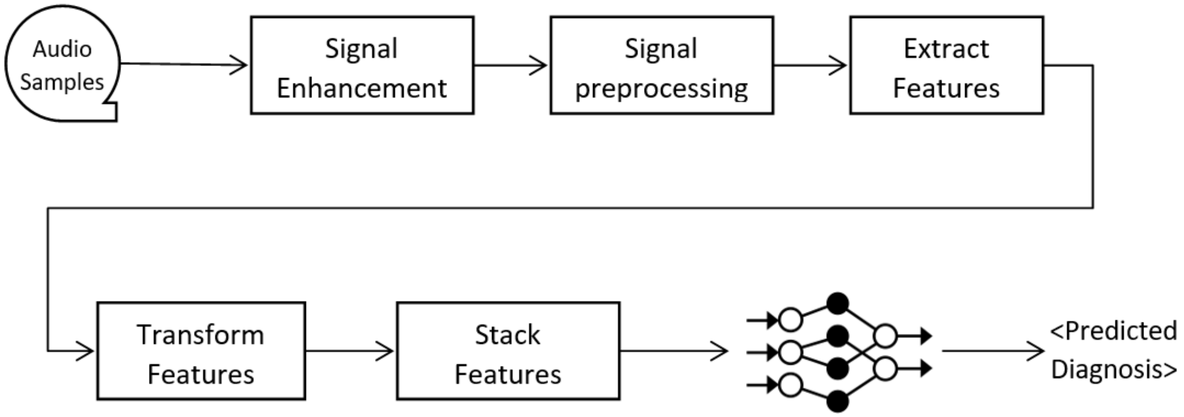


Figure 4: Process flow for classification of respiratory samples

The Dataset

The dataset used for this paper is the Mendeley Lung Sounds[2] dataset made available by King Abdullah University, Ramtha, Jordan. This dataset contains lung sound audio samples collected using an electronic stethoscope using a Bell, Diaphragm and extended filters. The Bell mode filter amplifies sound signals in the frequency range of 20-1000 Hz with emphasis on the low-frequency sounds between 20-200 Hz, the Diaphragm mode filter amplifies sounds in the frequency range of 20-2000 Hz with emphasis on frequency 100-500 Hz and the Extended filter which amplifies sounds in the frequency range 20-1000 Hz with emphasis on frequencies between 50-500 Hz range.

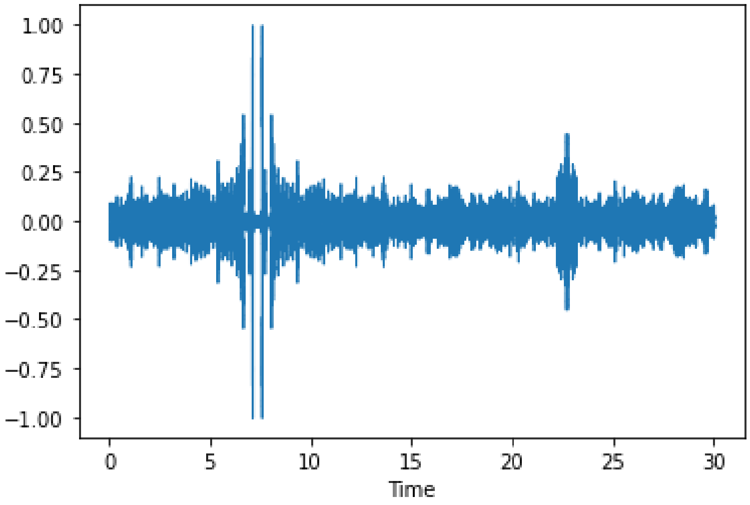


Figure 5: Waveform representation of lung sound sample

The dataset includes samples from 112 subjects, including 35 healthy and 77 patients between the ages of 12 to 90. The samples are recorded as ".wav" files and of a duration of between 4 and 30 seconds. The distribution of subjects based on their health conditions is as below:

Health Condition	No of Subjects	Age Range	Gender
Normal	35	18-81	11 female,24 male
Asthma	32	12-72	17 female, 15 male
Pneumonia	5	36-70	2 female, 3 male
COPD	9	42-76	1 female, 8 male
Bronchitis	3	20-68	1 female, 2 male
Heart failure	21	20-83	9 female, 12 male
Pleural Effusion	2	70-81	0 female, 2 male
Lung Fibrosis	5	44-90	2 female, 3 male

All samples are created using the sampling rate of 4000.

Table 1: Age and conditions of audio samples in the dataset[2]

Preparing the dataset

The following data preparation steps were taken to create a more balanced dataset, due to the discrepancies in the length and type of available samples:

1. Categorizing samples based on the diagnosis: Samples representing more than one diagnosis type (e.g. Lung fibrosis and heart failure) were replicated to separate samples for each condition.

2. Converting samples to equal duration: All samples were converted to an equal duration of 30 seconds by repeating the sequence and padding (reflecting the endings of samples).
3. Oversampling: Oversampling available samples to create a more balanced dataset with 35 samples in each class. This is performed by random selection from the available sample pool.

Extracting features

Audio data analysis requires the extraction of key features of the digitized sound for processing. An analog sound wave is digitized by quantizing its amplitude and frequency by sampling it at predefined discrete intervals known as the sampling rate. The higher the sampling rate the more closely the digitized samples will resemble the original sound wave. The downside of a higher sampling rate is that duplicate samples might be captured and a higher volume of samples resulting in larger files, therefore sampling rate is typically chosen so that the shape of the original waveform is preserved without losing too much of the detail.

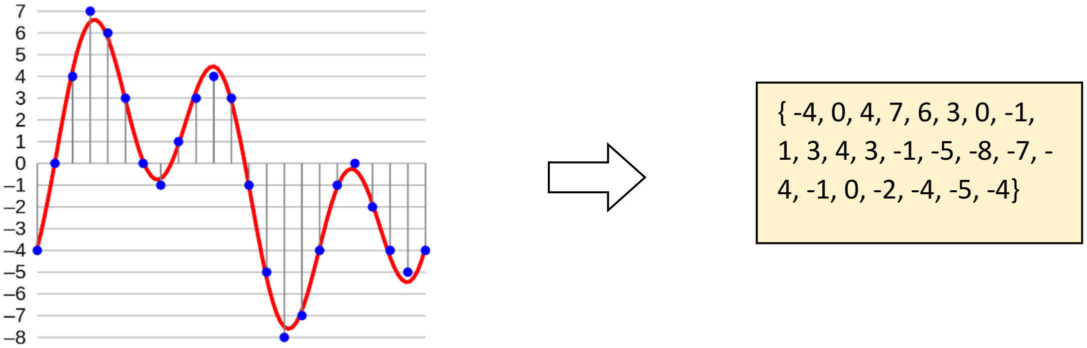


Figure 6: Digitized representation of an analog signal

Typically, an audio signal can be analyzed by extracting and evaluating its principle properties including amplitude and frequency but a more focused analysis also requires extraction of key features based on the nature of the signal and recording device used to reduce variability and noise. Feature extraction is performed by transforming the audio signal waveform to a parameterized representation with a lower data rate for subsequent processing and analysis. Since the lung sound samples are towards lower frequency bands, three separate audio feature extraction techniques have been used for analysis[4].

1. Short-Term Fourier Transform (STFT): The Fourier transformation model transforms signals between two different domains for example transforms a time-domain audio signal to its corresponding frequency domain. It demonstrates that a waveform can be represented by the sum of its constituent sinusoidal components. Each component is characterized by amplitude, frequency, and phase provides valuable insights into the nature of the original signal.

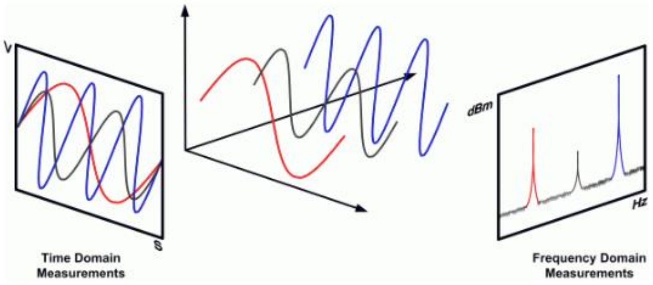


Figure 7: Visualization of Fourier transformation

Mathematically, the Fourier transform of a continuous signal f(t), can be represented as:

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt$$

Where *e* represents the base of the natural logarithm and *i* is the imaginary unit. An inverse Fourier transform function transforms a signal from the frequency domain back to the time domain. Although a powerful tool for signal analysis, the base Fourier transformation has some limitations. It is limited to analyzing signals that are continuous in time and cannot be used for signals that are discretely sampled at a constant interval and are limited in time or periodic. Furthermore, it cannot provide simultaneous time and frequency localizations, is computationally expensive and is sensitive to noise.

Short-term Fourier Transform (STFT) overcomes these challenges by breaking the signal into shorter segments rather than processing the entire signal at once. It works by applying the Fourier transform on each segment separately. STFT also provides a compromise between frequency and time domain representations by allowing information on frequency components of the signal at various points in time. STFT of a signal x at time t and frequency f can be defined as :

$$\mathrm{STFT}(t, f) = \int\text{-}\infty\wedge\infty\ x(\tau) * w(t\text{-}\tau) * e\text{-}j2\pi ft\ d\tau$$

where the segment length is defined by window function *w(t-τ)*, of typically a finite duration. The size of the window should be chosen carefully, typically a shorter window of enough length to capture signal fluctuation within the interval and a wider window for slowly fluctuating signals. There exists a trade-off between optimal time and frequency localization that can be achieved by the selected window length.

The output of STFT provides valuable insights into how the frequency content of a signal varies over time. The STFT results are converted to a spectrogram for analysis using a deep-learning convolutional neural network. A spectrogram provides a visual representation of a signal's properties with respect to time. Typically, waveform features such as frequency and amplitude are scaled and transformed into discrete intensity values and plotted as a heatmap with varying colors representing the strength of the signal property. STFT allows analysis of a signal's properties at a particular point in time. For this analysis. results of STFT are transformed to amplitude strength values for plotting over a spectrogram depicting amplitudes of different frequencies.

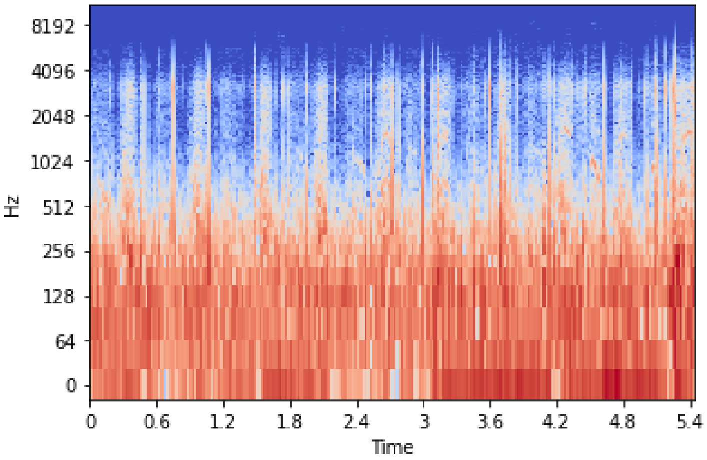


Figure 8: STFT Spectrogram of a lung sound sample

2. Mel-Scale Spectrogram: The human perception of sound is not linear in nature; we are more adept at perceiving variations in audio signals in the lower 20-500 Hz band than in the higher 10000 Hz or above range. The Mel scale[4] proposed by Stevens, Volkmann and Newmann proposed a unit of pitch so that frequencies at equal distances should be equally distinguishable. The Mel scale applies a logarithmic scaling on the audio signal such that sounds of equal distance on the Mel scale have the same perceptual distance. The Mel scale is nearly linear at lower frequencies and nearly logarithmic at higher frequencies.

$$m=2595\cdot\log(1+f/500)$$

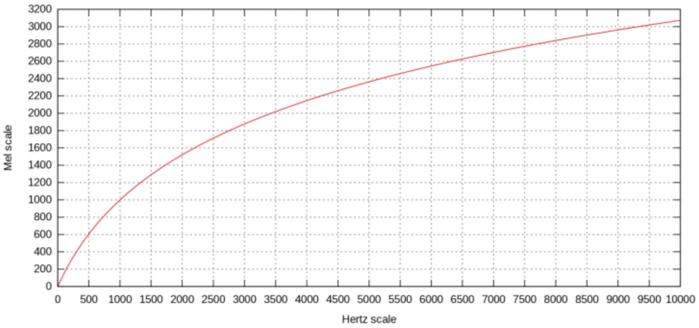


Figure 9: Visualization of the Mel-scale

Once frequencies of a signal are transformed using the Mel scale, they can be plotted to generate a Mel Spectrogram. A Mel spectrogram provides a time-frequency representation of the acoustic properties of a signal including frequency, energy and amplitude.

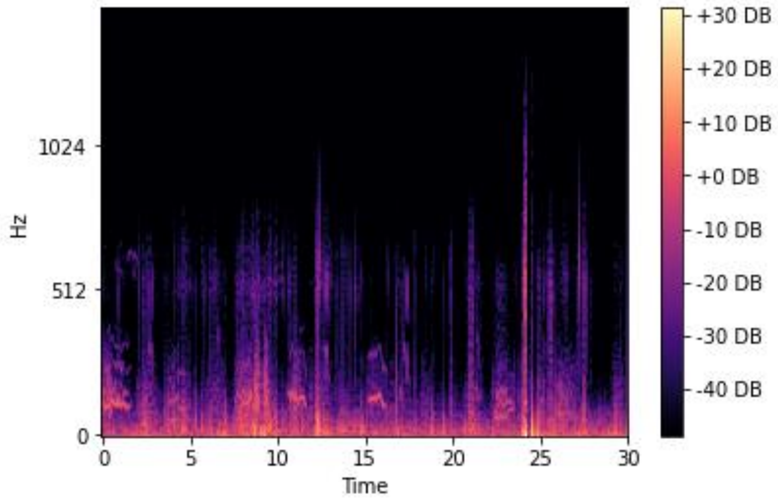


Figure 10: Mel-spectrogram representation of a lung sound sample

3. Mel-Frequency Cepstral Coefficients (MFCC): The Mel Frequency Cepstral Coefficients[4] (MFCC) is another way of extracting acoustic information from an audio sample using the Mel scale. Unlike the Mel-Spectrogram which provides a time-frequency representation, MFCCs are a set of coefficients that provide a compact representation of the spectral features of the audio signal by applying Discrete Cosine Transformation (DCT) over the Mel-Spectrogram of the signal.

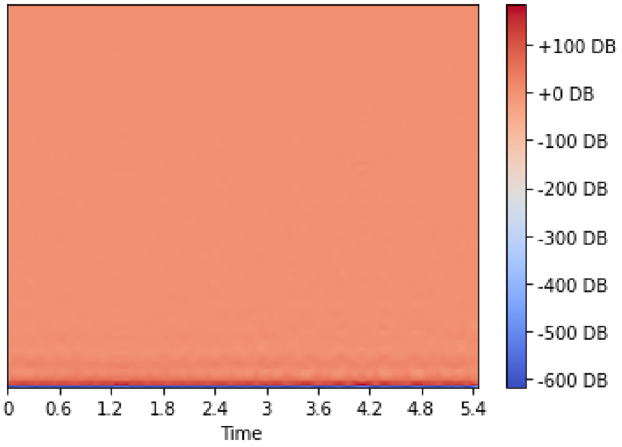


Figure 11: MFCC plot of a lung sound sample

Once all 3 features are extracted from a sound sample, they are stacked on top of each other to create a composite image with each feature represented as one channel of a 3-channel image. This composite 3D-image is used as the input for the convolutional neural network (CNN) for further analysis and classification.

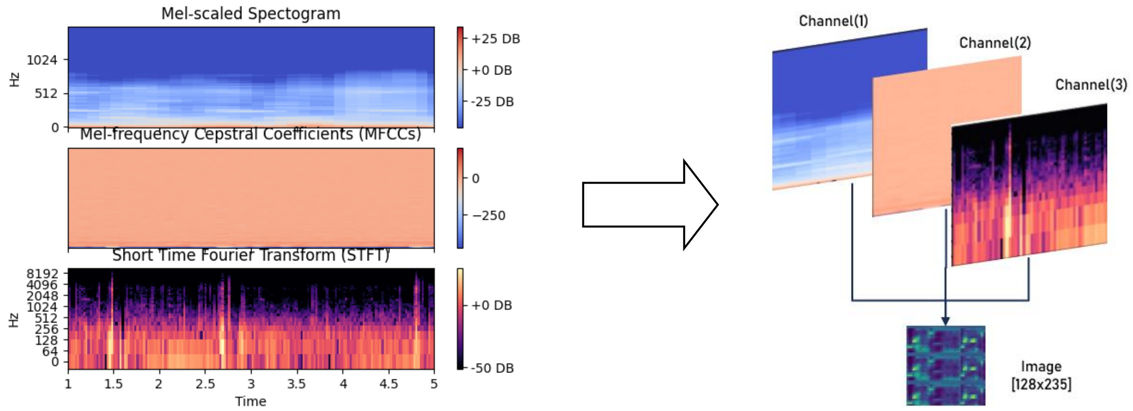


Figure 12: Creating 3D-image representation from extracted audio features

Baseline Conventional CNN Model

Two different deep neural network architectures were evaluated for the classification of lung audio samples. A custom 19-layer convolutional neural network was designed to provide a baseline evaluation of a conventional network architecture for analyzing audio features. The model uses 5 ‘convolution’ blocks of convolution, batch normalization and activation layers followed by a set of pooling, dense and dropout layers. The output layer of the model generates prediction scores for the 8 diagnosis classes represented by the dataset. An overview of the model architecture is shown below:

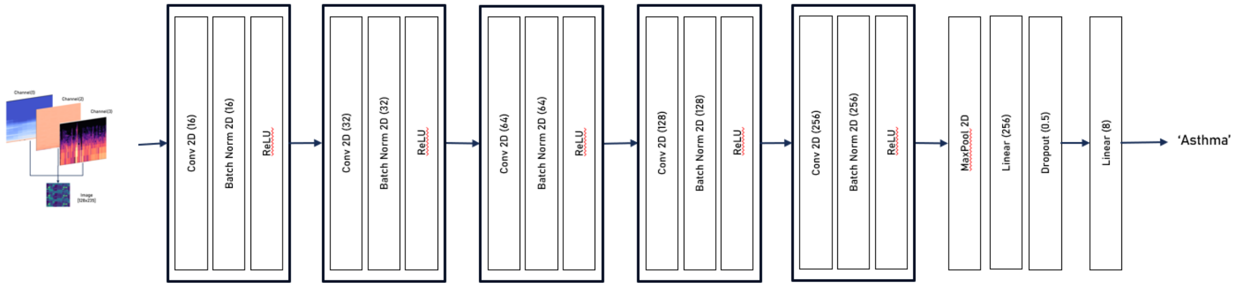


Figure 13: Model architecture of custom CNN Model

The architecture uses batch normalization (as described by Ioffe and Svegedy,2015) instead of dropout layers in the ‘convolution’ blocks to regularize results from the convolution layers and avoid model overfitting[6]. Batch normalization also provides resistance against internal covariate shifts while improving training time and convergence. Furthermore, considering the computing requirements of modern deep neural network architectures the use of dropout layers to facilitate regularization is inherently wasteful. Typically, the features processed by convolutional layers are highly correlated and require less regularization, making dropout layers less effective.

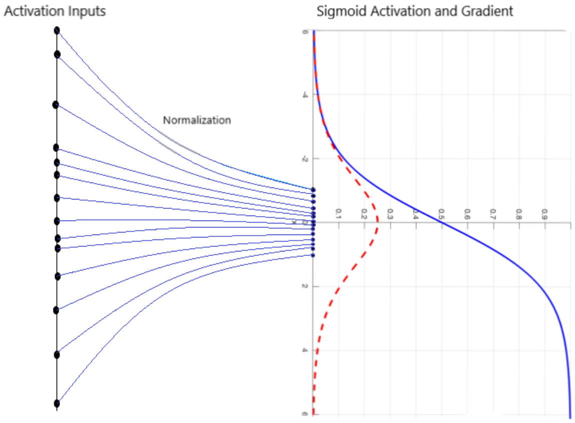
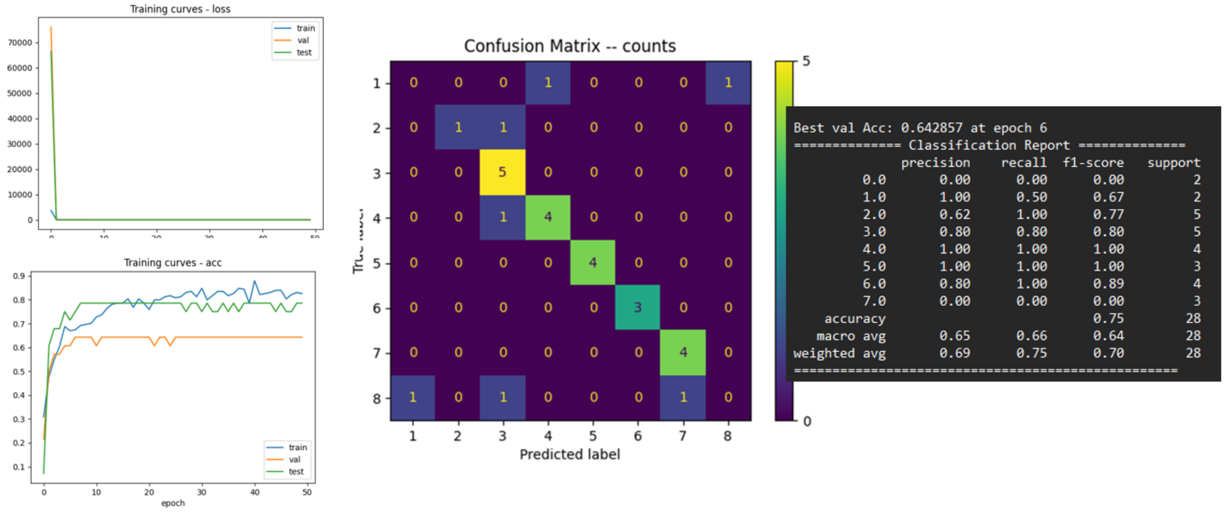


Figure 14: Batch normalization process

The network was trained using the 3D-image representation of the respiratory lung sound samples as inputs for 50 epochs. An exponential learning rate of 0.01 is used with the cross-entropy loss function. The results of the training cycles performed are below. The model was able to achieve an accuracy of 75%.



	1	2	3	4	5	6	7	8
1	0	0	0	1	0	0	0	1
2	0	1	1	0	0	0	0	0
3	0	0	5	0	0	0	0	0
4	0	0	1	4	0	0	0	0
5	0	0	0	0	4	0	0	0
6	0	0	0	0	0	3	0	0
7	0	0	0	0	0	0	4	0
8	1	0	1	0	0	0	1	0

```
Best val Acc: 0.642857 at epoch 6
============== Classification Report ==============
              precision    recall  f1-score   support
         0.0       0.00      0.00      0.00         2
         1.0       1.00      0.50      0.67         2
         2.0       0.62      1.00      0.77         5
         3.0       0.80      0.80      0.80         5
         4.0       1.00      1.00      1.00         4
         5.0       1.00      1.00      1.00         3
         6.0       0.80      1.00      0.89         4
         7.0       0.00      0.00      0.00         3
    accuracy                           0.75        28
   macro avg       0.65      0.66      0.64        28
weighted avg       0.69      0.75      0.70        28
==================================================
```

Figure 15: Training and validation results from baseline custom CNN model

Dense Convolution Network (DenseNet)

Dense convolutional network or DenseNet is a deep convolutional neural network architecture first introduced by Gao Huang et. al. in their paper Densely Connected Convolutional Networks, 2018[5]. Traditional convolutional neural networks typically rely on a deep chain of successive convolution blocks where each convolution block takes inputs from the convolution block just above it. Unfortunately, as the network grows in depth it becomes more and more difficult to pass the learnings from the lower layers to the upper layers via back propagation of gradient updates due to the vanishing gradient problem[9]. Due to the depth of the network, the delta updates to the gradients get smaller and smaller during back-propagation. Modern neural networks such as ResNet, HighwayNet and DesnseNet use several techniques including the use of residual connections to avoid the vanishing gradient issue. Residual or skip connections provide a shortcut by bypassing one or more intermediate convolution blocks in a deep convolutional neural network to directly feed in information to lower convolution layers from one or more upper layers across convolution blocks. This skipped information from the upper layers allows the model to learn from several sources and the application of batch normalization ensures that gradients do not vanish with depth.

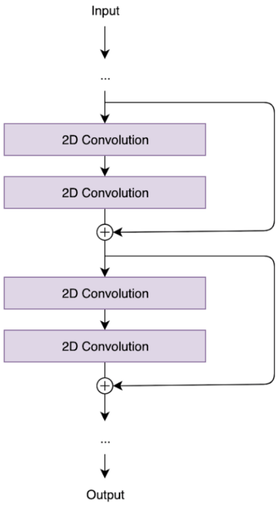


Figure 16: Skip connections typically feed inputs from previous layers

DenseNet takes the concept of residual connections to its extreme, where each layer of the network is directly connected to every other layer in the network thereby creating L(L+1)/2 dense connections for a L-layer network. At each layer input from the previous layers are concatenated to be used as inputs. This allows DenseNet to require fewer parameters and the ability to reuse feature maps derived earlier in the network with the Lth layer receiving feature maps from all previous layers. Unlike ResNet, in DenseNet feature states from the previous layer are not transformed by addition to the current input states but are preserved by concatenation before use, this dense connectivity feature of DenseNets make it well suited for complex domain-specific classification tasks such as medical image processing.

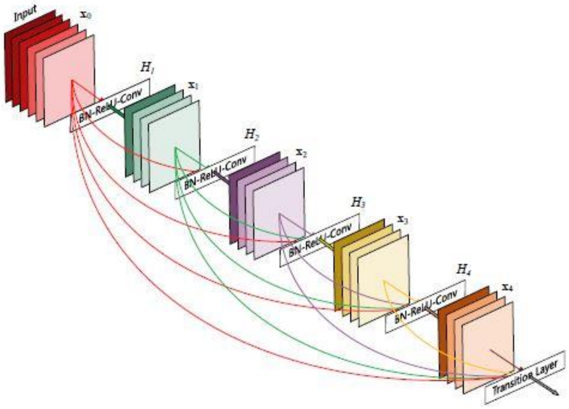


Figure 17: Feature map flow through the DenseNet

The concatenation action used by DenseNets is not possible if the size of the feature map changes as in the case of conventional CNNs due to sub-sampling. To achieve this the architecture of a DenseNet is divided into a series of dense blocks and transition blocks where the dimensions of a feature map remain unchanged within a dense block. Transition blocks are used for down sampling the feature maps before feeding into the next dense block.

For analysis of features extracted from the respiratory lung sounds, a version of DenseNet referred to as DenseNet-121[8] is used. The model architecture uses 1 7x7 Convolution Layer,58 3x3 Convolution Layers, 61 1x1 Convolution Layers, 4 Avg Pool Layers and 1 Fully Connected Layer. The semantic structure of this model is as shown below. The 3D-image representation of lung audio samples is fed into the network as input.

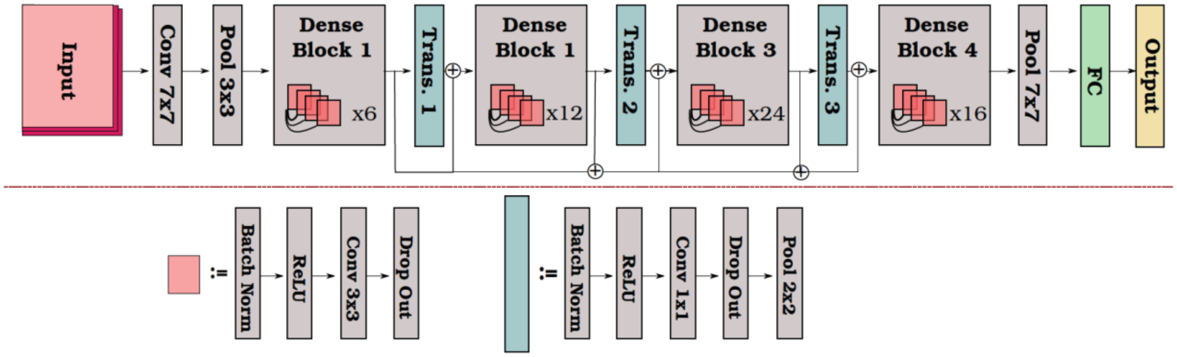
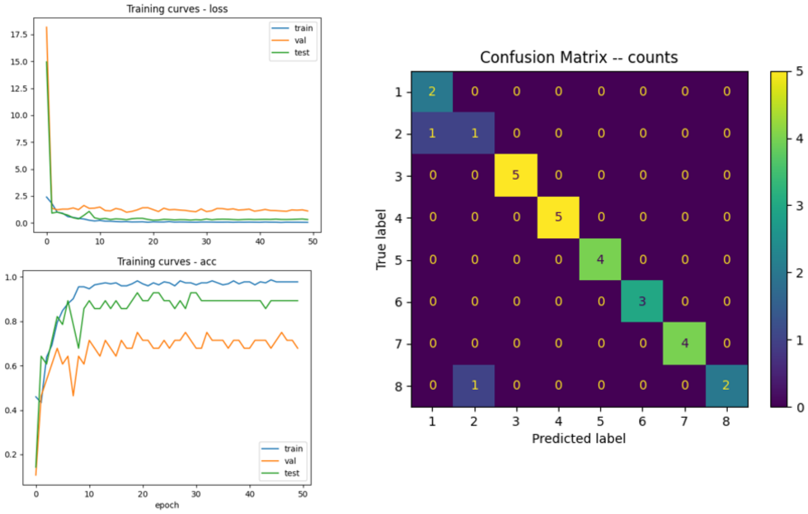
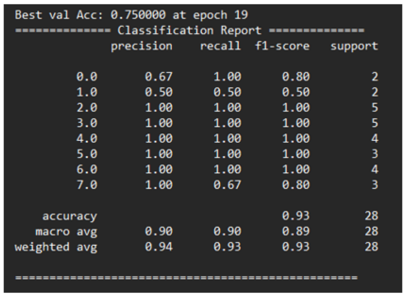


Figure 18: Semantic structure of DenseNet-121 Model[10]

The network is trained using the 3D-image representation of the respiratory lung sound samples as inputs for 50 epochs. An exponential learning rate of 0.01 is used with the cross-entropy loss function. Even with the low epoch count the network is able to achieve a respectable accuracy of up to 93% with all but two samples in the evaluation dataset being predicted correctly.





```
Best val Acc: 0.750000 at epoch 19
============== Classification Report ==============
              precision    recall  f1-score   support

         0.0       0.67      1.00      0.80         2
         1.0       0.50      0.50      0.50         2
         2.0       1.00      1.00      1.00         5
         3.0       1.00      1.00      1.00         5
         4.0       1.00      1.00      1.00         4
         5.0       1.00      1.00      1.00         3
         6.0       1.00      1.00      1.00         4
         7.0       1.00      0.67      0.80         3

    accuracy                           0.93        28
   macro avg       0.90      0.90      0.89        28
weighted avg       0.94      0.93      0.93        28

====================================================
```

Figure 19: Training and validation results from DenseNet CNN model

Conclusion

With this paper. I am attempt to demonstrate that analog signals in particular the sound samples can be analyzed using modern deep learning architectures. Audio feature extraction techniques such as Short-Term Fourier Transform, Mel-Spectrogram and Mel-Frequency Cepstral Coefficients (MFCC) can be applied to represent and transform key features from audio samples. The features can then be analyzed and classified using deep convolutional neural networks to provide accurate predictions which can help medical professionals in their diagnosis processes. The ability of such models to sift through thousands of samples in a short time can allow for better availability and coverage of medical help needed by patients with respiratory diseases.

References

1. Stephanie M. Levine, Darcy D. Marciniuk: Global Impact of Respiratory Disease (Global Impact of Respiratory Disease (chestnet.org))
2. Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, Ali Ibnian: A dataset of lung sounds recorded from the chest wall using an electronic stethoscope (A dataset of lung sounds recorded from the chest wall using an electronic stethoscope - Mendeley Data)
3. Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017 : (Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017 - The Lancet Respiratory Medicine)
4. Haytham Fayek : Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between (Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between | Haytham Fayek)
5. Gao Huang, Zhuang Liu, Laurens van der Maaten and Kilian Q. Weinberger: Densely Connected Convolutional Networks (1608.06993.pdf (arxiv.org))
6. Harrison Jansma : Don't Use Dropout in Convolutional Networks (Don't Use Dropout in Convolutional Networks - KDnuggets)
7. Tao Zhou ,XinYu Ye ,HuiLing Lu ,Xiaomin Zheng, Shi Qiu,5 and YunCan Liu : Dense Convolutional Network and Its Application in Medical Image Analysis (Dense Convolutional Network and Its Application in Medical Image Analysis (hindawi.com))
8. Aaliyah Ahmed: Architecture of DenseNet-121 (Architecture of DenseNet-121 (opengenus.org))
9. Yugesh Verma : Addressing The Vanishing Gradient Problem (Addressing The Vanishing Gradient Problem (analyticsindiamag.com))
10. Noha Radwan: Leveraging Sparse and Dense Features for Reliable State Estimation in Urban Environments ((PDF) Leveraging Sparse and Dense Features for Reliable State Estimation in Urban Environments (researchgate.net))

About the author

Kuljeet Singh, in a career spanning 20+ years, has worked extensively in the fields of Artificial Intelligence, Internet-of-Things, Multimedia and Embedded Systems domains. Presently serving as a Solutions Architect- IoT & AI with Wipro's Engineering Edge (WEE) division, he has designed and deployed several large AI solutions for worker safety, logistics, utility and manufacturing. He is also working with the technology teams of leading customers and industry experts on the next generation of IoT and AI solutions. For more information, contact Kuljeet at kuljeet.singh@wipro.com.