Using respiratory sounds to detect presence of Chronic Obstructive Pulmonary Diseases with Machine Learning

Aryaman Trivedi, Palaash Goel, Ieshaan Awasthy, Divyajeet Singh
Dept. of Computer Science and Engineering
Indraprastha Institute of Information Technology Delhi, India
{aryaman21022, palaash21547, ieshaan21054, divyajeet21529}@iiitd.ac.in

Abstract

Chronic Obstructive Pulmonary Diseases (COPDs) are a group of diseases that cause breathing-related problems and affect millions of people worldwide. COPDs are usually diagnosed by auscultation - the process of listening to organ sounds for diagnosis. However, this requires extensive experience and is prone to human error. In this paper, we propose a machine learning-based solution to detect COPDs using respiratory sounds. Our machine-learning model detects COPDs as accurately as the most experienced medical professionals.
*We use a dataset of 920 annotated audio samples collected from 126 subjects. This paper presents an analysis of models we tried to solve this problem. We achieved an accuracy score of **0.8967**+ using our top-performing model, exceeding the accuracy of most medical professionals.*

1. Motivation

Motivated to solve a real-world problem, the authors focused on the medical domain due to its profound health implications. Minor mistakes by medical professionals may manifest as life-lasting health effects. Exploring areas where such issues can be mitigated led us to the problem of detecting COPDs.

Detecting COPDs is a crucial task requiring extensive experience in auscultation - listening to the sounds of organs such as the heart and lungs to help find a medical diagnosis. However, Kim et al. [3] found a significant lack of accuracy in detecting COPDs among professionals. Most medical professionals, including resident doctors, detect COPDs with an average accuracy of only 61%, except fellows who exhibit an accuracy of 84%. This misdiagnosis of COPDs can lead to a significant increase in medical emergencies and a higher risk of mortality, as the patients' actual conditions remain untreated, as noted by Diab et al. [2].

To address this problem, we have provided an ML-based solution that can accurately detect COPDs using respiratory sounds. By doing so, we aim to improve the quality of care for patients with respiratory diseases and reduce the risk of misdiagnosis while reducing the stress on healthcare systems.

2. Introduction: Problem Statement

As mentioned previously, medical practitioners severely lack the ability to accurately diagnose chronic obstructive pulmonary diseases due to the skill and difficulty associated with auscultation. This challenge can be tackled with a machine learning-based solution that can reliably detect COPDs accurately using respiratory sounds. Medical professionals can use the solution to aid their diagnosis.

3. Literature Review

1. Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning by Kim et al.

   Studies the performance of medical professionals and CNNs in detecting CPODs from 1918 auscultation sounds collected from 831 patients. The authors split the audio into 6-second clips with 50% overlap and generated variations of Mel-Spectrograms as inputs for the CNN. They found that a frozen VGG16 trained on the ImageNet dataset (frozen transfer learning) for feature extraction, followed by a single convolution layer for classification, performed with an accuracy of 86.5%. They also found that medical students, interns, and residents struggle to identify CPODs, and only fellows could accurately detect CPODs from audio.

2. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease by Srivastava et al.

   Uses CNNs and DL methods with 10-fold cross-validation to provide a rigorous analysis of 920 annotated audio samples collected from 126 subjects for

COPD detection. The authors compared various features generated from the samples using the Librosa python library, including Mel-Spectrograms and Chromagrams. Their system also attempts to interpret the severity of the disease identified, such as mild, moderate, or acute, with an ICBHI score of 93%. The authors find that Mel-Frequency Cepstral Coefficients (MFCCs) provided better accuracy in detecting COPD than other features.

This study sheds light on some potential paths for us to take for our machine-learning models and feature extraction. Further, they provide domain knowledge of other authors and help us understand the data better.

4. Dataset

The Respiratory Sound Database provided by [1] is a collection of annotated audio samples collected independently over several years by two research teams from the the University of Coimbra & the University de Aveiro, Portugal, and the Aristotle University of Thessaloniki, Greece.

The dataset contains 920 samples annotated by respiratory experts recorded from 126 subjects spanning all age groups using heterogeneous recording devices. There are a total of 5.5 hours of clean and noisy recordings containing 6898 respiratory cycles - 1864 contain crackles, 886 contain wheezes, and 506 contain both.

Data Compilation

Despite its diversity, the size of the dataset is very limited, mainly when limited to numerical values from the 126 patients. Demographic data for these patients was extracted from a plain text file from the dataset.

Numerical Data

The 920 recordings have been taken from multiple microphones in different locations for varying durations. Most patients don't have recordings in every configuration, while others have recordings with duplicate configurations. The configurations for each recording were encoded within the filename. For example, `101_1b1_Al_sc_Meditron.wav` corresponds to patient number `101`, recording index `1b1`, taken in the Left Anterior (`Al`) location, in single channel mode (`sc`), using the Meditro Stethoscope. The annotations for each respiratory cycle (cycle_beginning_time, cycle_ending_time, crackles_present, wheezes_present) were contained in the same folder with the same name as the recording in plain text format, as shown below:

```
0.036 0.579 0 0
0.579 2.450 0 0
2.450 3.893 0 0
```

We attempted to create a single numerical dataset by compiling each patient's data into a percentage value for each chest location. Models trained on this dataset are treated as baselines for our future models.

A disadvantage to this approach was temporal invariance - by compressing all the breathing cycles into a singular number (representing the proportion of cycles where crackling or wheezing was detected, averaged across multiple recordings), we lost potentially relevant data about ordering these cycles. We also lost the single/multi-channel data in the process, although the loss should not affect our numerical predictions.

Audio Data

To address the disadvantages, we utilize four techniques. A set of features from audio files was generated using Mel Frequency Spectrograms and MFCCs, which capture the power distribution of frequencies in the signal for each phenome of a sound. Being scaled on the mel frequency scale allows them to accurately mimic human perceptions of sound. Mel Spectrograms carry many of the same advantages as MFCCs, but provide a detailed representation of the frequency content of an audio signal and provide sharper resolution.

More feature maps were generated, specifically using Chromagrams, representing energy distribution across the twelve pitch classes in Western music, and Chroma-Energy Normalized Statistics (CENS), which smooth out local deviations in audio features such as pitch and articulation.

This way, we end up with four different feature maps to train models on. These techniques of audio-feature extraction have been studied extensively in [4].

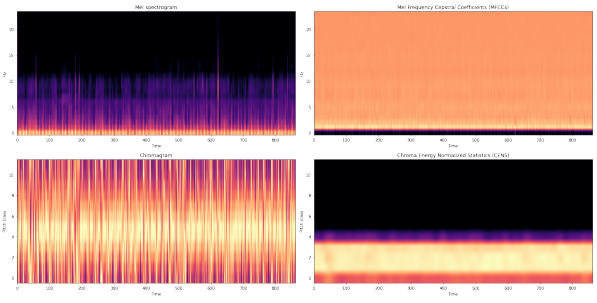


Figure 1. Discriminatory Features Extracted from Audio Files

Preprocessing

Audios were clipped or padded to 20 seconds each. We perform augmentation on the dataset to increase the robustness of the trained model, deal with the problem of our limited dataset, and artificially generate more (and varied) training samples. The techniques of time shifting, pitch shifting,

time stretching, and adding random noise were used for augmentation. The preprocessing introduces some degree of invariances with respect to time and pitch in the extracted features.

Data Analysis and Visualization

An exploratory data analysis determined features relevant to training an optimal model. Firstly, the distribution of the demographics revealed that most data points in the dataset corresponded to adults, most of whom were unhealthy. Hence, it became apparent that to scale up the project, separate models may be needed to predict the presence of COPDs in children and adults.

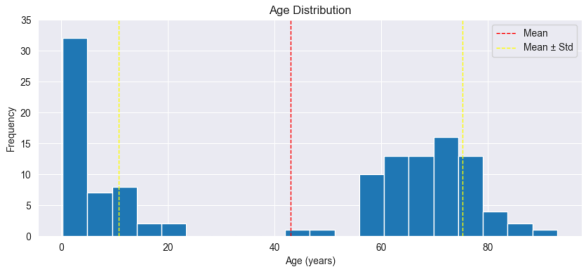


Figure 2. Distribution of Age

In fact, BMI is not defined for children, and the heights and weights of children were given in the dataset.

The distribution of the diseases present in the database strongly suggests that the database is imbalanced. Most adults were found to be unhealthy - most of them had general COPD, while the distribution was roughly equal for children.

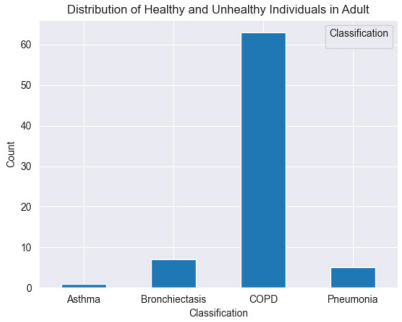


Figure 3. Distribution of specific COPDs in Adults

Finally, a distribution of the percentage of cracking and wheezing in respiration was drawn, classifying the patients as healthy and unhealthy. As expected, the highest mean of both was displayed by unhealthy adults. This clearly indicates that these sounds are good discriminating features, at least to classify whether a person is healthy.

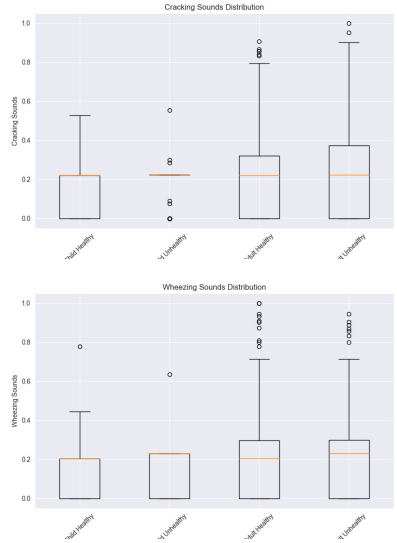


Figure 4. Distribution of Cracking and Wheezing Sounds

5. Methodology

The dataset provided sets of information in two forms - numerical and audio. All relevant information can be captured with the audio data. Various machine learning models were trained using the numerical data and audio data separately to classify patients into eight categories, including healthy and seven specific COPDs.

Models on Numerical Data

The classification based on the numerical data features extracted from the dataset served as baselines. Specifically, we showed that using only the numerical data, the Support Vector Machine model with Linear Kernel was above to achieve an overall accuracy of ≈ 0.8 in the case of adults and children. Models trained on the numerical data enabled us to analyse the efficacy of using the presence of wheezing and crackling in a patient as markers of different COPDs. Since wheezing and crackling have been recorded as features in the given dataset, it is imperative to explore how 'difficult' it is to diagnose a person based on these features. Two distinct datasets were created for prediction: one for adults and another for children. This division was necessary because adult BMI data was unavailable for children, and conversely, child height and weight information was missing for adults. Consequently, separate classifiers were trained for each subset. Due to these variations in available features, creating a single model that could uniformly handle all data points wasn't feasible.

The accuracy for each model was validated using k-fold cross-validation with different values of k, including 3, 5, and 10. Even though we plan to analyse the models more intensely to find the scope for improvement, we believe that

given a larger dataset, even such numerical features will produce a *good enough* model. To verify the classifier's quality obtained, we also look at the models' precision, recall, and F1-score.

Models on Audio Data

The audio pipeline contained the following steps:

1. *Preprocessing*: This included padding/truncating raw audio files to 20s of length.

2. *Augmentation*: To expand the audio dataset and make the models more robust to new data, augmentation was performed. This included:

    - Random Noise: Addition of random noise to the audio files.
    - Time Stretching: Stretching or compressing the audio files and padding them to 20s of length.
    - Time Shifting: Moving the waveform in time, and wrapping it around the 20s window.
    - Picth Shifting: Increasing/decreasing the pitch of the file.

    The augmentation was performed randomly on all the files to create a new augmented dataset of size 920*5=4600 samples.

3. *Feature Extraction*: This consisted of two steps:

    (a) Audio Feature Extraction: For extracting data from audio, the following were generated from each audio file:

    - Mel Frequency Spectrogram
    - Mel Frequency Cepstral Coefficients (MFCCs)
    - Chromagrams
    - Chroma Energy Normalised Statistics (CENS)

    These allowed us to represent the audio in a 2D numerical array, making them machine readable.

    (b) Numerical Feature Extraction The audio features generated could be passed through a CNN to generate final set of numerical features by flattening its output, or simply flattened directly into a large-dimensional array.

4. *Classification* Classification was performed via the CNN by using the flattened features it output as inputs to an ANN. Meanwhile, the directly flattened features from the audio features were classified using the following models:

    - SVM
    - KNN
    - Random Forest
    - Logistic Regression
    - Naive Bayes

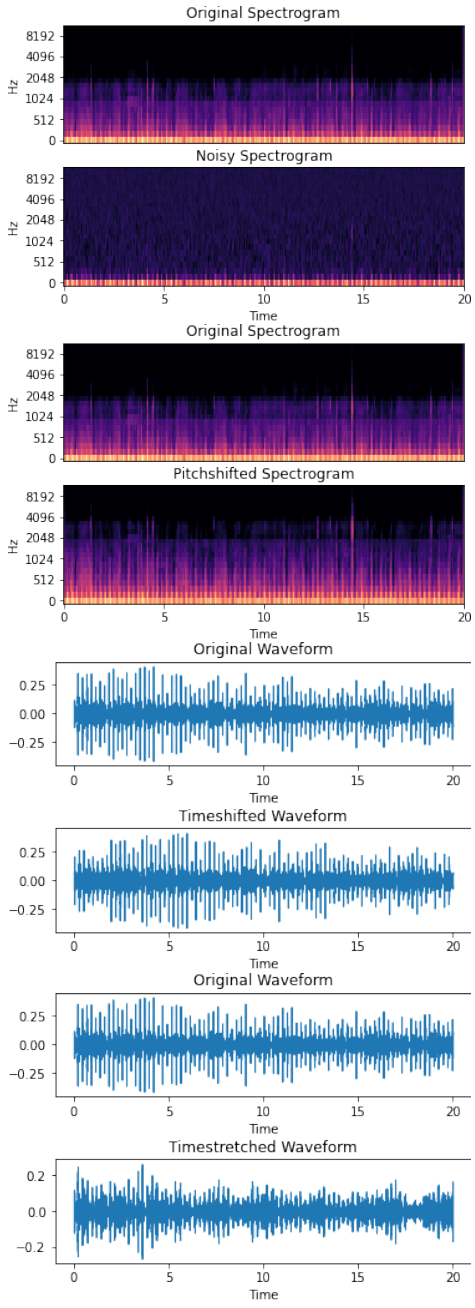
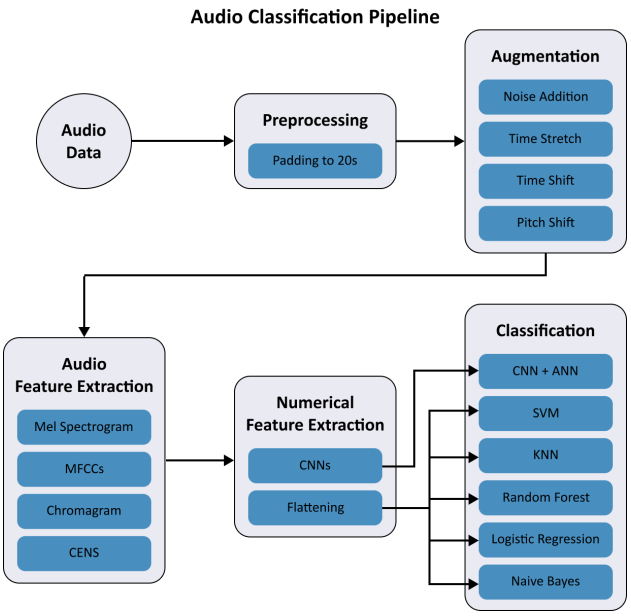


Figure 5. Augmented vs Original Audio for Patient 101

In order to train machine learning models to learn generalised patterns in the respiratory audio data and classify these sounds into one of eight categories, the first step required feature extraction from this audio. These features



were extracted in the form of Mel Spectrograms, Mel-Frequency Cepstral Coefficients (MFCC) Spectrograms, Chromograms, and Chroma Energy Normalised Statistics (CENS), all of which condense the audio information into visual plots. These visual plots effectively converted the audio classification problem into an image classification problem for which we used multiple CNN-based models - one model for each type of image plot. The idea was to compare the discriminating power of these feature extraction methods and identify which method allowed the model to learn generalised patterns in the dataset most effectively.
All of these CNNs had a similar architecture - consisting sequentially of a convolution layer, a max-pooling layer, and another convolution layer, all three of which together perform the task of feature extraction from images followed by flattening the multi-dimensional representation matrix for each feature into a single vector. These layers were succeeded by three linear layers, which learn patterns from these flattened representations. Some parameters, such as the input and output sizes of the linear layers, were experimented with to cater to a specific type of feature set.

A range of classifiers were tried for the problem on both the augmented and non-augmented flattened datasets, including SVMs, k-NN, and random forests.

6. Results and Analysis

Models on Numerical Data

The lineup of models trained on numerical data included techniques from Gaussian Naive Bayes to Decision Trees and SVMs. A typical pattern is observed among all models - the prediction accuracy of children is generally lower than that of adults. This can be attributed to the fact that there are more adult training samples. Also, adults in the dataset have been classified into fewer classes (there are classes into which no adults have been classified). Similarly, fewer data points of children are classified into a larger number of classes - this is a consequence of having limited training data.

The performance of the models is tabulated in Table 1.

S.No.	Model	Training Accuracy		Testing Accuracy	
		Adult	Child	Adult	Child
1	*Support Vector Machines*	**0.807**	**0.55**	**0.875**	**0.6**
2	*k-Nearest Neighbors*	0.836	0.403	0.75	0.4
3	*Decision Trees*	0.752	0.4	0.625	0.6
4	*Decision Trees*	0.721	0.649	0.75	0.6
5	*(Gaussian) Naive Bayes*	0.77	0.516	0.5	0.6
6	*Logistic Regression*	0.838	0.514	0.875	0.6

Table 1. Training and Testing Accuracies for Different Models

The success of the logistic regression and SVM-based models can be used to infer the linear separability of the given data. This is further supported by the fact that the linear kernel function performed the best among the different kernel functions used for the SVM classifier. As previously stated, the SVM model with the linear kernel was the top performer. The slightly worse performance of logistic regression can be attributed to the increased robustness of SVMs to outliers.

Models on Audio Data

As anticipated, extracting features directly from the respiratory audio and then performing multi-class classification on them yielded much better results than in the case of numerical data. These results have been tabulated in Table 2. Models were created for the augmented audio datasets

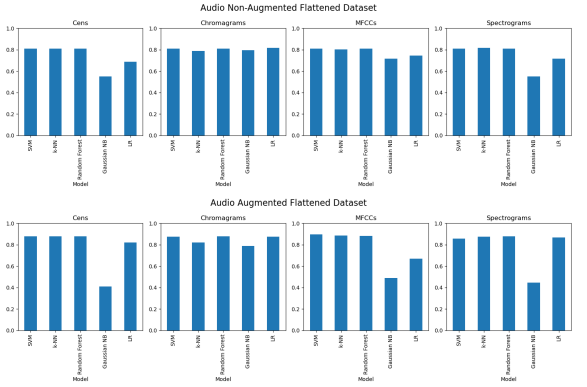
S.No.	Model	Accuracy		
		Training	Validation	Testing
1	*CNN on Spectrograms*	**0.874**	0.815	0.815
2	*CNN on MFCC*	0.855	**0.902**	**0.880**
3	*CNN on Chromagram*	0.865	0.848	0.848
4	*CNN on CENS*	0.870	0.826	0.837

Table 2. Training and Testing Accuracies for Different CNN Models

as well. However, the high computational requirements for training these models combined with our lack of resources to meet them, we were unable to actually train these models and evaluate the difference in performance brought on by augmenting the original dataset. It is important to note that according to our literature review, we anticipate an increase in performance due to augmentation.

The performance of other models trained on both augmented and non-augmented data can be seen as follows.

Performance: Flattened Audio Features						
S.No.		1	2	3	4	5
Model		SVM	k-NN	RF	Gaussian NB	Log. Reg.
Base Dataset	CENS	*0.8116*	0.8116	0.8116	0.5507	0.6884
	Chromagrams	0.8116	0.7899	0.8116	**0.7971**	*0.8188*
	MFCCs	*0.8116*	0.8043	0.8116	0.7174	0.7464
	Spectrograms	0.8116	*0.8188*	0.8116	0.5507	0.7174
Augmented Dataset	CENS	*0.8786*	0.8786	0.8786	0.4112	0.8225
	Chromagrams	0.8768	0.8225	*0.8804*	0.7899	**0.875**
	MFCCs	***0.8967***	**0.8877**	**0.8841**	0.4909	0.6703
	Spectrograms	0.8587	0.8768	*0.8786*	0.4457	0.8696



Among the numerical models, SVMs performed the best consistently, but failed to perform well for children, likely due to the limited data. On the audio dataset, the CNNs demonstrated the greatest scope for future models, with models trained on the base dataset generally matching the performance of models trained on the augmented flattened data. The flattened models also demonstrated that augmentation improved the quality of predictions across the range significantly.

7. Conclusion

We have created a model achieving an accuracy of 89.67%, exceeding the accuracy of the most experienced medical professionals (81%). This score was achieved by the SVM model trained on the augmented dataset with the MFCC-generated features. Hence, we have improved upon our SVM-benchmark model which achieved a base score of around 0.8, trained on the numerical data.
The MFCCs generally generated the best predictive features for classifiers.

Ultimately, we hope the model will enable more effective treatment and improve patient outcomes, using only basic numerical and audio data.

References

[1] L. Mendes I. Vogiatzis E. Petrantoni E. Kaimakamis P. Natsiavas A. Olivera C. Jácome A. Marques R.P. Paiva I. Chouvarda P. Carvalho N. Maglaveras B.M. Rocha, D. Filos. A respiratory sound database for the development of automated classification, 2018.

[2] Nermin Diab. Underdiagnosis and overdiagnosis of chronic obstructive pulmonary disease, 2018.

[3] Yoonjoo Kim. Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning, 2021.

[4] Miranda R Patil S Pandya S Kotecha K. Srivastava A, Jain S. Deep learning-based respiratory sound analysis for detection of chronic obstructive pulmonary disease, 2021.