Multi-Respiratory Diseases Classifications:
Implementing and Comparing Different Machine Learning Techniques

Lihan Feng & Ivana Li

1. Abstract

Respiratory diseases such as COVID-19 have brought many impacts on our lives. It is urgently needed to have a cost-effective technique to support current auscultation techniques to accurately diagnose diseases. Audio signal changes caused by respiratory disease bring hope in using machine learning techniques to classify diseases based on these signal changes. In this project, we implement both traditional machine learning techniques such as decision tree and support vector machine (SVM) as well as deep learning models such as VGG16 and AlexNet to classify 8 different respiratory diseases based on 1676 different respiratory cycles labeled with 8 different diseases. Audio features are converted to MFCC 2D arrays and Mel-spectrogram RGB figures for traditional and deep learning models. Using VGG16 architecture and training all the parameters to extract features from Mel-spectrogram and using a multi-layer perceptron classifier gives the best accuracy of classifying the disease (95.2381%). Our study shows that it is promising to use CNN techniques to study respiratory characteristics by converting audio signals to Mel-spectrogram. However, domain shift problems should be noted while using a pre-trained model. Future improvements of this project include detailed noise removal of such as heartbeat sounds. It is also important to notice the imbalance distribution of different diseases in real life while creating validation and test sets for the model training process.

2. Introduction and Background

Respiratory disease is an important category of disease that greatly influences people's health status. This category includes both chronic diseases such as asthma, cystic fibrosis, and chronic obstructive pulmonary disease (COPD) and communicable diseases such as COVID-19 and influenza. With high morbidity and prevalence, it is important to diagnose the disease accurately and efficiently. Auscultation has been a cost-effective way to diagnose respiratory disease at clinical use. However, accurate interpretations of various respiratory sounds require clinicians' sufficient knowledge about the field and lead to the low efficiency of diagnosing due to the limitations of expert clinicians. [1]

Recently diagnoses based on AI techniques have come to the stage because the nature of inflammation caused by various diseases can lead to audible changes which can further be identified as diagnostic signals. Therefore, many studies have implemented machine learning techniques to study the audio signals of patients with respiratory diseases. Over the past few

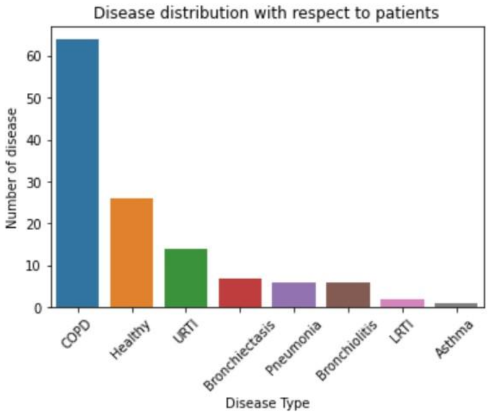
[1] Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning | Kim et al.

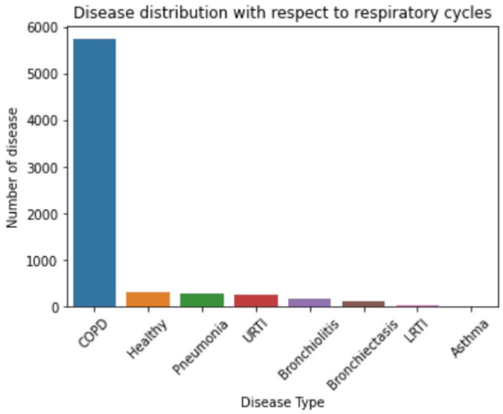
years, there is a study that categorized respiratory sounds into normal sounds, sounds with crackles or wheezes, and so on by using a deep learning convolutional neural network (CNN) in the clinical setting. [2] There are also studies using computer audition to do speech and sound analysis to automatically recognize and monitor the symptoms of COVID-19. [3]

3. Description of problem

Our task is to use machine learning models to classify whether someone has respiratory diseases and what kind of diseases someone may have based on their audio data of respiratory cycles. Furthermore, we compare the performance of different machine learning models.

We use the dataset originally from the "Respiratory Sound Database" in Kaggle (link). This dataset contains 6898 respiratory cycles from 126 people. Respiratory sounds were recorded in various chest locations with multiple equipments. Detailed information about the acquisition of the audio can be found in the file name of each audio and its corresponding annotation text file.





4. Methodology

4.1 Data processing

4.1.1 Text-file preprocessing & Data imbalance handling

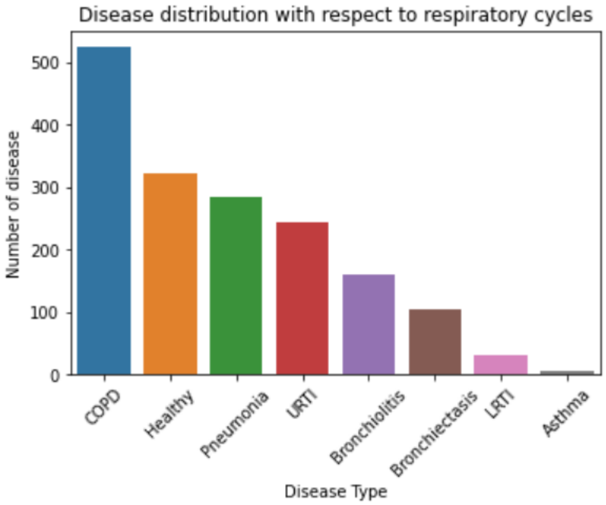
We first process text files and merge patient ID, the corresponding labels of respiratory diseases, corresponding audio file names, and the timestamp of each cycle in each audio file into a data frame. Then we observe the basic information of our raw data.

Based on the chart of the distribution of raw data, we face one of the most common issues in classification tasks with observation data, data imbalance. With some classes having scarce data compared to others, the models can extract limited information from the training set, which causes inaccuracy. Therefore, we dropped some patients' data with the label "COPD" (Chronic

[2] Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning | Kim et al.

[3] COVID-19 and Computer Audition: An Overview on What Speech & Sound Analysis Could Contribute in the SARS-CoV-2 Corona Crisis | Schuller et al.

obstructive pulmonary disease) to make the data balanced. We can observe the new distribution of preprocessed data in the following figure.



4.1.2 Audio data extraction

In the dataset, one audio file contains several respiratory cycles for one patient, we first need to extract audio data. To make sure our input data has the same dimension, we need to extract the respiratory cycles under the same duration. Based on the Box plot of duration (figure 1), we select 5s as the length to maintain the information. For cycles duration above 5s, we directly trim it down and for those under 5s, we add zero padding (add silence) at the end. We use packages librosa and soundfile to realize the process of audio data extraction.

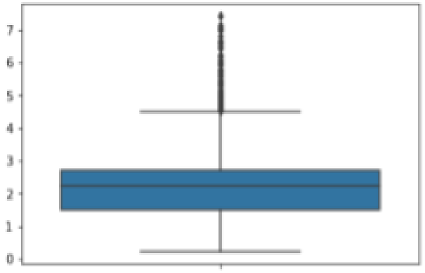


Figure 1: Box plot of the duration

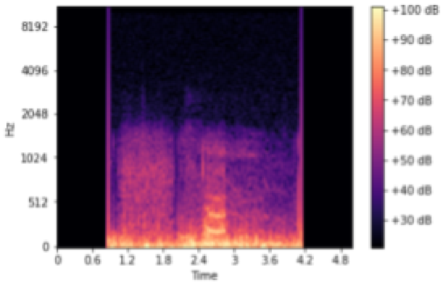


Figure 2: Mel Spectrogram Sample

4.1.3 Mel Spectrogram creation

For an audio classification problem, one way to solve it is to transform it into an image classification problem. Therefore, we transfer our audio data into one type of image, Mel Spectrogram. A spectrogram is a visualization of the frequency spectrum of a signal, where the frequency spectrum of a signal is the frequency range that is contained by the signal. Mel spectrogram is a spectrogram converted into a Mel scale[4]. Mel scale is a non-linear transformation of the frequency scale which mimics how the human ear works[5]. Since the Mel

[4] Detect covid19 with CNN: Detect covid-19 from Mel Spectrogram | Analytics Vidhya | Goni, A.

[5] Mosaic: A classical machine learning multi-classifier based approach against Deep Learning classifiers for Embedded Sound Classification | Lhoest et al.

spectrogram is a representation of audio data and its transformation method is similar to humans, we choose to convert audio data into Mel Spectrogram. We realize the conversion through package librosa.

4.1.4 MFCC conversion

MFCC stands for Mel Frequency Cepstral Coefficients, which is widely used in audio classification, whose computation is based on a Mel frequency scale, which is consistent with Mel Spectrogram[6]. Besides, the return value of MFCC for one audio file is an array, which can be the input for the classifiers in traditional machine learning. Therefore, we can compare the performance of different deep learning models and traditional machine learning models under the same criterion. We write a convert_to_mfcc() function based on package librosa to realize the conversion.

4.2 Model Creation & Evaluation

4.2.1 Traditional Machine Learning Classification Technique

Traditional Machine Learning based models generally consist of two stages: 1) Extract the acoustic features from audio signals 2) Create and train a classifier to predict disease classes based on the acoustic features. In practice, there is a wide range of frequently used acoustic features within two classes: temporal features such as tempo, period, and beat-loudness as well as spectral features including MFCC. [7] In our case, we choose spectral features MFCC, which is the frequency-domain feature after Fourier transformation, as our input for the traditional machine learning pipeline.

We implement two frequently used classification models Decision Tree and Support Vector Machine (SVM) with one using a tree-based classification technique and the other one as the linear separator.

4.2.2 Deep learning Image Classification Technique

Convolutional Neural Network is a widely used technique in order to extract distinguishing features from 2D figures. Therefore, we used two popular CNN models, AlexNet and VGG16. The two models are compared with pretrained parameters and with random initialization. When using a pre-trained model, we freeze the feature extractor part of the CNN model while training the last few classifier layers. When we use random initialization, we essentially train everything from scratch. Then we apply a multi-layer perceptron to realize the classification task by adding another layer to that output 8 logits for our 8 different classes. For the input of CNN, we use Mel-spectrogram images where RGB information represents amplitudes and different locations in width and height represent different times and frequencies.

[6] Mosaic: A classical machine learning multi-classifier based approach against Deep Learning classifiers for Embedded Sound Classification | Lhoest et al.

[7] Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues | Xia et al.

AlexNet is a deep convolutional neural network with 60 million parameters with 5 convolutional layers, 3 max-pooling layers, 2 normalization layers, 2 fully connected layers, and 1 softmax layer. It performs well in recognizing off-centered objects. It paved the way for the future development of other deep models such as VGGNet and ResNet and so on.

VGG16 is also a popular model for image classification tasks, which is able to classify 1000 images of 1000 different categories with 92.7% accuracy. The architecture of VGG16 consists of 21 layers with 16 layers having trainable parameters. VGG16 is special in having consistent small-size 3x3 kernels in the convolution layers that help extract detailed information about the image.

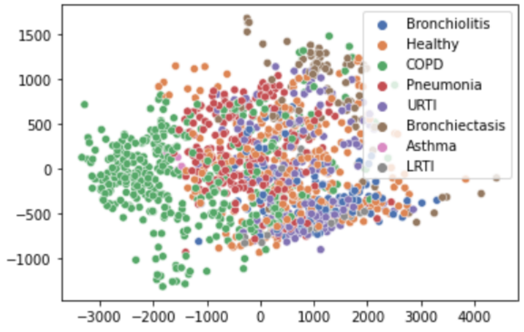
5. Results

5.1 Traditional Machine Learning Technique

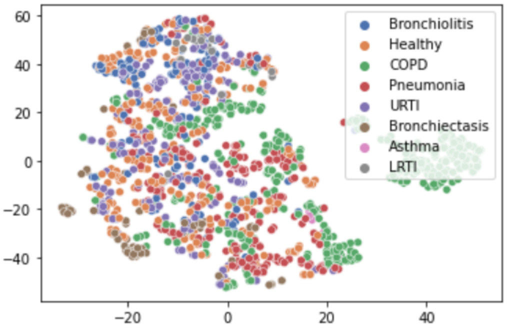
	Decision Tree	SVM
Accuracy	59.5238%	74.7023%

We found out that SVM using the linear kernel performs better than the Decision Tree. This result may be due to the reason that SVM generates a linear classifier that performs better in our dataset. In particular, we use linear and non-linear kernels in the SVM and found out that the linear kernel performs the best while the other kernels perform far from satisfaction (sigmoid kernel: 29.1666%, RBF kernel: 31.25%, polynomial kernel of degree 2: 48.2143%). This further shows that our MFCCs features might be more linear separable.

In the below figures, we use dimensionality reduction techniques PCA and t-SNE to visualize our input, high-dimensional MFCC vectors. PCA reduces dimension by preserving the variance while t-SNE focuses on preserving the high-dimensional relative distance. We can observe that some classes, such as the most frequent class COPD, can possibly be separated by a linear separator in the high dimensional space.



PCA



t-SNE

5.2 Deep learning Image Classification Technique

	VGG 16 Pre-trained	VGG 16 Not Pre-trained	AlexNet Pre-trained	AlexNet Not Pre-trained
Accuracy	57.4405%	95.2381%	64.5833%	91.6667%

We found out that 1) Not pre-trained VGG16 performs better than not pre-trained AlexNet. 2) Not pre-trained model performs much better than the pre-trained model.

First of all, VGG16 can generally be seen as an improved version of AlexNet. VGG16 has more convolutional layers. It also uses a smaller size filter to capture the detailed information in the image. In Mel-spectrogram, it is important for the model to be able to recognize the change of amplitude in terms of different frequencies to distinguish different figures. Therefore, VGG16 might be a better choice compared to AlexNet. However, we can still observe that in transfer learning, AlexNet performs better than VGG16. This phenomenon is because the dataset that VGG16 previously trained on might deviate more from our image dataset compared to AlexNet's pre-trained dataset.

We analyze that the reason for the second finding is that trainable parameters in AlexNet and VGG16 are previously trained and fixed on the ImageNet dataset. ImageNet consists of human-annotated figures such as animals. However, in our case, our images for classification are Mel-spectrograms which is not an image of a typical, regular-shaped object. Therefore, it is better to train all the trainable parameters based on our dataset instead of only training the classifier layers. The problem can be seen as domain shifts where the dataset for the pre-trained model and the dataset for our classification have significantly different distributions and characteristics.

Finally, by comparing the traditional machine learning and deep learning techniques, we found that feature extraction using CNN can greatly improve classification accuracy. This is reasonable because CNN is able to read the image data directly and analyze it while maintaining the 2D characteristics. On the other hand, when we directly flatten the MFCC 2D array and feed it into traditional machine-learning models, the 2D structure is no longer preserved. .

6. Limitations

6.1 Noise removal

In the real world, the audio samples we collected often have low SNR (signal-to-noise ratio). The higher SNR is, the higher proportion of signals accounts for. In our case, though the accuracy of deep learning models like VGG16 or AlexNet is high, we didn’t do noise removal before we trained the models, which makes the performance become relatively worse.

6.2 Problems with the validation set

The strategy we chose to deal with data imbalance is to directly drop some patients' data. This strategy is equivalent to generating a new dataset from the original dataset to train models. Since the strategy was used before we split the training set and validation set, the validation set contains balanced data. However, based on information from the original dataset, the data in real life is imbalanced. Therefore, the validation set we gave to the model has a gap to the situation in real life. Though we used to consider using methods like SMOTE to automatically generate more data of classes with fewer samples, it’s complex to implement this to audio data, and potentially it will lose some information related to time or frequencies.

7. Future improvements

Studies suggested that re-sampling audio recordings to 4 kHz and deploying a fifth-order Butterworth band-pass filter having 100–200 Hz cut-off frequencies can effectively eliminate environmental noise such as heartbeat, motion artifacts, and audio sounds[8]. Therefore, we can first do de-noising before we transform them into Mel Spectrogram, which can further improve the performance of our models and reduce interferences like different kinds of noise.

8. Conclusions

In this project, we implement a classification task that takes audio files as inputs and classifies 8 different respiratory diseases. We use Mel Frequency Cepstral Coefficients (MFCC) that represents audio signals in the frequency domain as features for traditional machine learning models decision tree and SVM. We use RGB Mel-spectrogram as input figures for deep learning models, in particular AlexNet and VGG16, and classify the input using a Multi-layer perceptron. It turns out that training a VGG16 model all from scratch gives the best accuracy (95.2381%).

We find that deep learning models perform better than traditional machine learning models for preserving the 2D information of the image of the audio. Linear classifier SVM performs better than non-linear classifiers, which indicates the linear separable characteristics of our input dataset. Training all the trainable parameters in the pre-train model architecture can lead to better performance because of the domain shift issue. In particular, our input images are Mel-spectrograms that represent signals rather than common objects.

Nonetheless, there are some limitations in our project, particularly in noise removal and creating a realistic validation set that needs future improvements to build a more accurate and valid model to actually bring more AI-based diagnosis techniques into real life.

[8] Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues | Xia et al.

References

Goni, A. (2021, July 1). Detect covid19 with CNN: Detect covid-19 from Mel Spectrogram. Analytics Vidhya. Retrieved December 22, 2022, from https://www.analyticsvidhya.com/blog/2021/06/how-to-detect-covid19-cough-from-mel-spectrogram-using-convolutional-neural-network/

Kim, Y., Hyon, Y. K., Jung, S. S., Lee, S., Yoo, G., Chung, C., &amp; Ha, T. (2021). Respiratory sound classification for Crackles, wheezes, and Rhonchi in the clinical field using Deep Learning. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-96724-7

Leitner, B.Z., & Thornton, S. (2019). Audio Recognition using Mel Spectrograms and Convolution Neural Networks.

Lhoest, L., Lamrini, M., Vandendriessche, J., Wouters, N., da Silva, B., Chkouri, M. Y., &amp; Touhafi, A. (2021). Mosaic: A classical machine learning multi-classifier based approach against Deep Learning classifiers for Embedded Sound Classification. Applied Sciences, 11(18), 8394. https://doi.org/10.3390/app11188394

Schuller, B. W., Schuller, D. M., Qian, K., Liu, J., Zheng, H., &amp; Li, X. (2021). Covid-19 and Computer Audition: An overview on what Speech &amp; Sound Analysis could contribute in the SARS-COV-2 corona crisis. Frontiers in Digital Health, 3. https://doi.org/10.3389/fdgth.2021.564906

Xia, T., Han, J., &amp; Mascolo, C. (2022). Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. Experimental Biology and Medicine, 153537022211154. https://doi.org/10.1177/15353702221115428

Zhou, Q., Shan, J., Ding, W., Wang, C., Yuan, S., Sun, F., Li, H., &amp; Fang, B. (2021). Cough recognition based on Mel-spectrogram and Convolutional Neural Network. Frontiers in Robotics and AI, 8. https://doi.org/10.3389/frobt.2021.580080