Classification of Lung Sounds using Machine Learning

Xiaoran Xu
iCONS Lab, Dept of Electrical Engineering
University of South Florida
Tampa, Florida, USA
(xiaoranxu@usf.edu)

In-Ho Ra
School of Computer, Information and Communication Engineering
Kunsan National University
Gunsan, South Korea
(ihra@kunsan.ac.kr)

Ravi Sankar
iCONS Lab, Dept of Electrical Engineering
University of South Florida
Tampa, Florida, USA
(sankar@usf.edu)

ABSTRACT

Lung sounds are physiological signals produced during the ventilation process between the human respiratory system and the outside world and have high research value for the analysis and diagnosis of different lung diseases. In particular, the Covid-19 pandemic in recent years has greatly increased the need for rapidity and accuracy in the diagnosis. Lung auscultation has attracted widespread attention due to its convenience and non-invasiveness in the development of automatic lung sound diagnosis technology. The development of hardware such as an electronic stethoscope and other signal acquisition technologies has further promoted the research and progress of modern lung sound signal analysis and technology. In this study, the lung sound dataset provided by Shanghai Children's Medical Center (SCMC) was classified using the combined method of MFCC-CNN. In the event-level classification, the binary classification problem achieved an accuracy of 0.81, while the multi-classification achieved an accuracy of 0.80. Both the record-level binary classification and the multi-classification achieved an accuracy of 0.69.

KEYWORDS

Lung Sound, Machine Learning, MFCC, CNN

1 INTRODUCTION

The lung sound signal is a physiological sound signal produced by the human respiratory system and the outside world in the process of ventilation. The mechanism is complex and contains rich physiological and pathological information [1]. Macroscopically, lung sounds can be divided into normal and abnormal breath sounds. Abnormal breath sounds are divided into Coarse Crackle, Fine Crackle, Rhonchi, Stridor, Wheeze, and Wheeze and Crackle.

Different lung diseases can be diagnosed by the detection of corresponding abnormal lung sounds, such as wheeze can be used to detect asthma [2], Rhonchi is effective for the diagnosis of chronic obstructive pulmonary disease [3], Crackle is associated with pneumonia and pulmonary fibrosis [4].

In recent years, due to the ongoing pandemic of Covid19, the need for fast and accurate diagnosis of lung diseases has exponentially increased. With the expanding lung sound research, the benefits and prospects of auscultation, a convenient and safe diagnostic method, in the diagnosis and treatment of lung diseases are becoming more and more realizable. However, in diagnosis, it is difficult to meet the growing demand for rapid diagnosis of lung diseases only by human auscultation by auscultation physicians. It can be more quantitative and automated. With the continuous development of machine learning algorithms and a large number of applications in practice, the idea of machine learning, which has been proven to be an effective technology, is constantly extending to more fields. Computerized lung sound analysis has also attracted the attention of many researchers in recent years [5]. Computerized lung sound signal processing based on machine learning technology is undoubtedly a more advanced field in lung sound research, which deserves more attention. At the same time, an electronic stethoscope provides a lot of help for us to collect lung sound signals. The automatic pattern recognition-based system can be embedded in an electronic stethoscope to cope with the limitations of traditional auscultation techniques.

In order to improve the diagnostic efficiency, we hope to use machine learning, as a preliminary diagnosis, to help us identify different lung sounds. Machine learning is the core of artificial intelligence and the fundamental way to make computers intelligent. It mainly uses inductive and comprehensive methods to acquire new knowledge and reorganize the existing knowledge structure to continuously improve the performance of the artificial intelligence system [6]. For lung sound signal processing, researchers expect that the constructed lung sound recognition classifier can continuously improve its judgment ability in the processing and discrimination of a large amount of lung sound data under appropriate algorithms and models, and finally achieve accurate effective automatic identification and classification. This challenge includes event-level binary classification (Normal and Adventitious) and multiclass classification challenges (Normal (N), Rhonchi (R), Wheeze (W), Stridor (S), Coarse Crackle (CC), Fine Crackle (FC), Wheeze and Crackle (WC)). Use sensitivity, specificity, average score, and harmonic score as reference performance indicators.

The main contribution of this paper is to illustrate the concept of using the Mel-scale Frequency Cepstral Coefficients (MFCC) and convolutional neural network (CNN) framework for binary and multi-classification of the lung sound dataset. Then, validate the classification of data collected by electronic stethoscopes to help doctors better distinguish between different types of lung sounds.

The organization of the paper as follows: Section II systematically summarizes the related work and Section III describes the dataset.

The description of the method is provided in Section IV followed by the conclusion and the direction of future research.

2 RELATED WORK

The following research work provide valid proofs of using various methods of feature extraction and classification of lung sounds.

Chowdhury et al. [7] used a combination of MFCC and linear predictive coding (LPC) to extract the feature values of sound information and use a 1D convolutional neural network for classification. This method is robust to audio degradation but fails to demonstrate differences between different audio signal datasets.

Ashar et al. [8] used the MFCC-CNN method to classify speakers and obtained an accuracy of 87.5%. The disadvantage is that the data set was relatively small, and other data sets were not used to compare and verify the effectiveness of the model. A CNN trained on MFCC features for diagnosing infant asphyxia by Zabidi et al. [9]. A single convolutional CNN was trained using the MFCC features of normal and asphyxiated infant cry signals. The results show that the method achieved a high accuracy rate, which proves that the method is very suitable for the diagnosis of neonatal asphyxia based on the non-invasive data acquisition method.

The following research work provide proofs of the effectiveness of the CNN model on lung sounds.

Aykanat et al. [10] tried to use the CNN algorithm in audio classification. Since MFCC features are combined with SVM as a benchmark. The spectrogram image classification using the CNN algorithm was found to be as effective as the SVM system. The CNN and SVM algorithms were run comparatively to classify breathing audio: rales, articulations, and normal voice classification, and the experimental accuracy results were 76% for CNN and 75% for SVM. Bardou et al. [11] used convolutional neural networks to classify lung sounds by using three types of inputs: spectrogram, MFCC features, and LBP features and obtained 95.56% accuracy using CNN. Tariq et al. [12] extracted three unique features from audio samples, namely Spectrogram, MFCC, and Chromagram. Finally, a fusion of the three best convolutional neural network models is constructed by inputting image feature vectors converted from audio features. The highest accuracy achieved was 99.1% based on Spectrogram lung sound classification but no multi-classification was reported.

The following research contributed to the classification of abnormal sounds in the lungs.

Faustino et al. [13] detected Crackle and wheeze based on CNN-MFCC and achieved results consistent with the current state-of-the-art, with an accuracy of 43% and a sensitivity of 51%.

3 DATASET

The dataset was gathered by Shanghai Children's Medical Center (SCMC) using a Yunting model II stethoscope [14] and stored in wav format. The overall audio length is 15.36 seconds, containing 1949 audio files and comments. Normal, CAS, DAS, CAS, and DAS are the annotations at the record level. The annotation at the event level consists of the start (ms) and finish (ms) of respiratory events, as well as the respiratory event type (Normal, Rhonchi, Wheeze, Stridor, Coarse Crackle, Fine Crackle, Wheeze and Crackle).

4 METHODS

We combine the Mel-scale Frequency Cepstral Coefficients (MFCC) and convolutional neural network (CNN) in this experiment.

4.1 MFCC

The MFCC is a model of human sound perception based on inner ear frequency analysis. The MFCC set provides perceptually meaningful, smooth estimates of the speech spectrum over time. How the structure of the human inner ear works: Mechanical vibrations generate a standing wave at the entrance of the cochlea, causing the basilar membrane to vibrate at maximum amplitude at this frequency at a frequency commensurate with the frequency of the incoming sound wave.

For audio, the classic MFCC extraction processing is divided into pre-emphasis, framing, windowing, fast Fourier transform (FFT), Mel filter bank filtering, logarithm, and discrete cosine transform (DCT) as shown in Fig. 1.

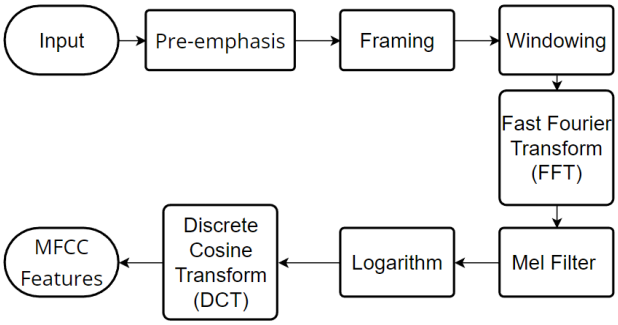

Figure 1: MFCC feature extraction processing.

The process of extracting MFCC in Python librose does not have a pre-emphasis step but directly performs framing. If not specified, the frame length defaults to 2048 samples, and the frameshift defaults to 512 samples. For the matrix obtained after framing, the next step is to add a window (LibROSA is first windowed and then framed). The purpose of windowing is to eliminate the discontinuity between frames after framing to a certain extent (the default Hanning window is used). After processing in the previous step, the obtained result is subjected to a fast Fourier transform (FFT) frame by frame. The process of fast Fourier transform frame by frame is called short-time Fourier transform (short-time Fourier transform or short-term Fourier transform, STFT). After the short-time Fourier transform, the absolute value needs to be taken, and then the energy spectrum can be obtained after squaring.

While obtaining the energy spectrum of the audio, it is also necessary to construct a Mel filter bank and perform a dot product operation with the energy spectrum. The role of the mel filter is to convert the energy spectrum into Mel frequencies that are closer to the human ear. For a small sound, the human ear can feel it as long

as the loudness is slightly increased, but when the loudness of the sound has reached a certain level, even if there is a greater increase, the human ear's feeling does not change significantly. We call this auditory characteristic of the human ear's loudness of sound the "logarithmic" characteristic. Therefore, the reason for taking the logarithm of the Mel spectrogram is to simulate the "logarithmic" nature of the human ear. The last step is the Discrete Cosine Transform (DCT), the purpose of this step is to change the data distribution and separate redundant data. After transformation, most of the signal data will be concentrated in the low frequency region, so we usually only need to take the first part of the transformed data (LibROSA's mfcc function takes the first 20 samples by default, and this article uses 40 samples).

4.2 CNN

After the eigenvalues are extracted, the eigenvalues are put into the convolutional neural network as input for classification. The structure of the CNN model is as shown in Fig. 2. This time, 4 Conv2D volume base layers are used, and the filter size is 2x2. The input shape is (40, 646, 1), where 40 is the number of MFCC features and 646 is the number of frames. Each convolutional layer is followed by a max pooling layer, and each pooling layer is followed by a dropout layer with a value of 0.1. Following the final convolutional layer is a global average pooling layer, and the final section is a three-layer dense layer with the same number of neurons as the classification.

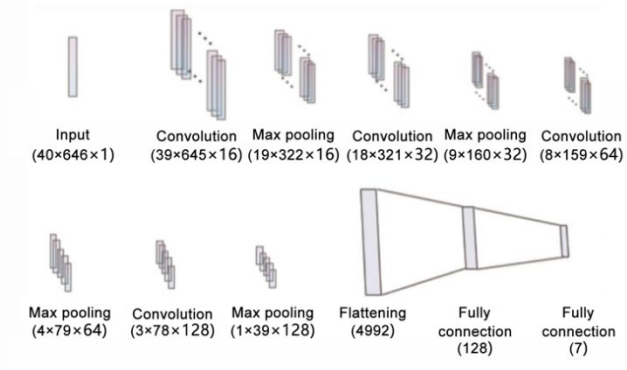

Figure 2: The structure of CNN.

The role of each convolution kernel is equivalent to a filter, the first convolution layer is 16 convolution kernels, the second layer is 32 and so on. In this way, expanding the yield channel enables the size of the bounds of the comparison between the two convolutional layers. The window state of the two max-pooling layers of the convolutional layer block is 2×2, and the step is 2. Since the shape of the pool window is similar to the step, the area covered by each sliding area of the pool window on the information does not cover each other. The pooling layer mainly has the following five functions: Increase the network receptive field, suppress noise, reduce information redundancy, reduce the amount of model calculation, reduce the difficulty of network optimization, and prevent network overfitting.

As the yields of the convolutional layer blocks are passed into the fully associative layer blocks, fully associative layer squares will smooth each example in a small-scale cluster. In summary, the information state of a fully associative layer is obtained in two dimensions, with the first measure being a fairly constant instance, the subsequent measurements being vector characterizations after each instance is flattened, and the vector length being the vector length. Channel, stature, and width results are the three fully associative layers in the fully associative layer block.

5 RESULTS

We use Sensitivity (SE), Specificity (SP), Average Score (AS), Harmonic Score (HS) as the main evaluation of the test results. Additionally, we show the confusion matrix and ROC curve. Dichotomous sensitivity was 0.25, specificity was 0.96, AS was 0.605, and HS was 0.39. Multi-class sensitivity was 0.56, specificity was 0.97, AS was 0.765, and HS was 0.71 (see Table 1). The advantage of this method is that the specificity is high, so the probability of successfully ruling out the disease in a population without the disease. The sensitivity performance is mediocre, indicating that the probability of successfully confirming the disease is not very good in the patient population.

Table 1: Evaluation Metrics

	Sensitivity	Specificity	Average Score	Harmonic Score	Accuracy
Record Level Binary Class	0	1	0.5	0	0.69
Record Level Multiclass	0	1	0.5	0	0.69
Event Level Binary Class	0.25	0.96	0.605	0.39	0.81
Event Level Multiclass	0.56	0.97	0.765	0.71	0.80

In Fig. 3 and Fig. 4, the ROC curve is a more intuitive display of our main evaluation. Among them, wheeze has the best performance in the ROC curve with an area of 0.91, and stridor has the worst performance with an area of 0.52. We can find that the classification accuracy of normal lung sounds is relatively high, but there are many misclassifications in the other abnormal lung sounds. As shown in Fig. 5 and Fig. 6, we observed that the most abnormal sounds are misclassified as normal. Abnormal sounds other than wheeze have not been correctly classified. The MFCC+CNN method was tested by event level and record level experiments using ensemble learning models. It was determined to be superior to continuous/discontinuous adventitious respiratory sounds in classifying normal and abnormal lung sounds. More methods will be tried in the future to improve the accuracy.

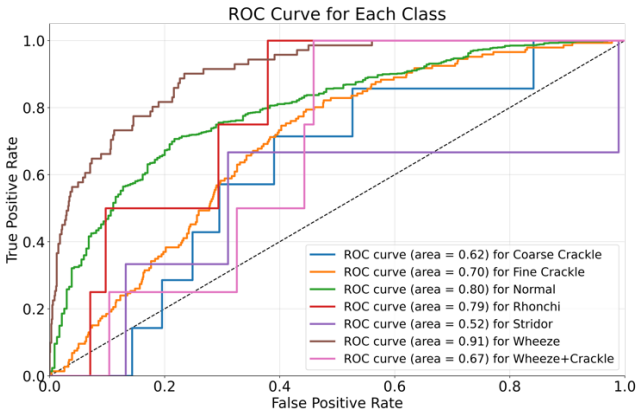

Figure 3: Event level ROC curve of multiclass.

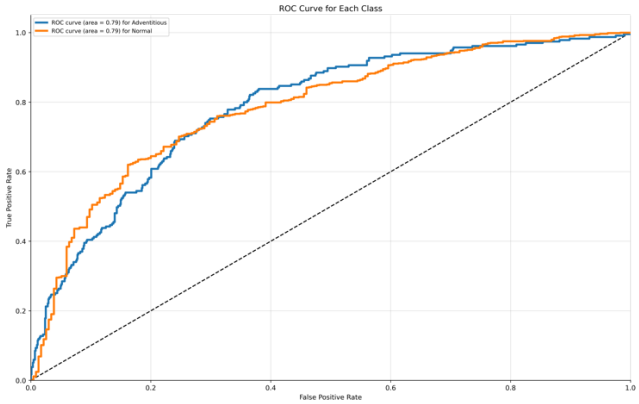

Figure 4: Event level ROC curve of binary class.

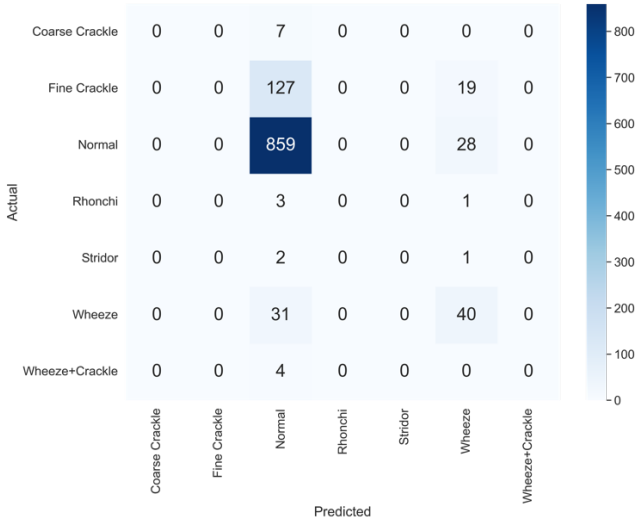

Figure 5: Event level confusion matrix of multiclass.

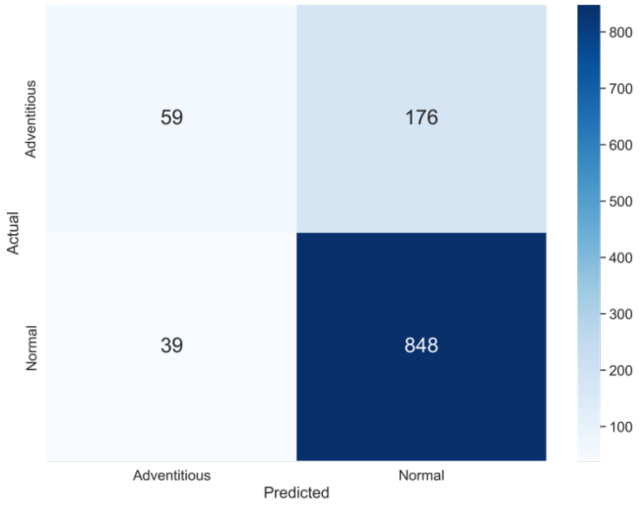

Figure 6: Event level confusion matrix of binary class.

6 CONCLUSION

In this work, we proposed classification of normal and abnormal lung sounds using the MFCC+CNN method and ensemble learning model for better diagnosis of respiratory diseases by lung sounds. The high specificity of the results obtained by this method indicates that there is less chance of misdiagnosis in the diagnosis of a disease and can be a good aid to the physician as a preliminary diagnosis. We show that our model has a good classification outcome in distinguishing abnormal lung sounds from normal lung sounds with an acceptable accuracy of dichotomous classification and multiclassification. It can help to successfully classify the diseased from the non-diseased population. It is our hope that our model can be helpful for pulmonologists and medical professionals/clinicians in treating respiratory diseases.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2014333).

REFERENCES

[1] Fraiwan, Luay, et al. "Automatic identification of respiratory diseases from stethoscopic lung sound signals using ensemble classifiers." Biocybernetics and Biomedical Engineering 41.1 (2021): 1-14.

[2] Shaharum, Syamimi M., K. Sundaraj, and Rajkumar Palaniappan. "Tracheal sound reliability for wheeze data collection method: A review." 2012 IEEE International Conference on Control System, Computing and Engineering. IEEE, 2012.

[3] Morillo, Daniel Sánchez, et al. "Computerized analysis of respiratory sounds during COPD exacerbations." Computers in biology and medicine 43.7 (2013): 914-921.

[4] Piirilä, Päivi, et al. "Crackles in patients with fibrosing alveolitis, bronchiectasis, COPD, and heart failure." Chest 99.5 (1991): 1076-1083.
Palaniappan, Rajkumar, Kenneth Sundaraj, and Nizam Uddin Ahamed. "Machine learning in lung sound analysis: a systematic review." Biocybernetics and Biomedical Engineering 33.3 (2013): 129-135.

[5] Palaniappan, Rajkumar, Kenneth Sundaraj, and Nizam Uddin Ahamed. "Machine learning in lung sound analysis: a systematic review." Biocybernetics and Biomedical Engineering 33.3 (2013): 129-135.

[6] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017.
[7] Chowdhury, Anurag, and Arun Ross. "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals." IEEE transactions on information forensics and security 15 (2019): 1616-1629.
[8] Ashar, Aweem, Muhammad Shahid Bhatti, and Usama Mushtaq. "Speaker identification using a hybrid CNN-mfcc approach." 2020 International Conference on Emerging Trends in Smart Technologies (ICETST). IEEE, 2020.
[9] Zabidi, A., et al. "Detection of asphyxia in infants using deep learning convolutional neural network (CNN) trained on Mel frequency cepstrum coefficient (MFCC) features extracted from cry sounds." Journal of Fundamental and Applied Sciences 9.3S (2017): 768-778.
[10] Aykanat, Murat, et al. "Classification of lung sounds using convolutional neural networks." EURASIP Journal on Image and Video Processing 2017.1 (2017): 1-9.
[11] Bardou, Dalal, Kun Zhang, and Sayed Mohammad Ahmad. "Lung sounds classification using convolutional neural networks." Artificial intelligence in medicine 88 (2018): 58-69.
[12] Tariq, Zeenat, Sayed Khushal Shah, and Yugyung Lee. "Feature-based Fusion using CNN for Lung and Heart Sound Classification." Sensors 22.4 (2022): 1521.
[13] Faustino, Pedro, Jorge Oliveira, and Miguel Coimbra. "Crackle and wheeze detection in lung sound signals using convolutional neural networks." 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021.
[14] Qing Zhang, "SPRSound: Open-Source SJTU Paediatric Respiratory Sound Database," [Online]. Available: https://github.com/SJTU-YONGFU-RESEARCH-GRP/Lung-Sound-Database/