Proceedings of the 2023 International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/28/20230401

Using upsampling CONV-LSTM with metadata embedding for respiratory sound classification

Changhe Chen[1, 2], Rongbo Zhang[1]

[1] Faculty of Applied Science & Engineering, University of Toronto, Toronto, Ontario, Canada, M5S 1A1.

[2] changhe.chen@mail.utoronto.ca

Abstract. Respiratory diseases are one of the leading causes of death around the world and they severely affect patient quality of life. Auscultation is an essential method for diagnosing respiratory diseases, and it is low-cost and convenient. However, auscultation requires experts who are highly experienced. Medical trainees suffer from misdiagnosis inevitably. To address this issue, a novel machine learning model is proposed, which consists of upsampling convolutional neural network (CNN), a long short-term memory network (LSTM), and a fully connected network (FCNN) with embedding layers to classify respiratory sounds into seven categories: Normal (N), Rhonchi (R), Wheeze (W), Stridor (S), Coarse Crackle (CC), Fine Crackle (FC), Wheeze & Crackle (WC). The model is trained and evaluated on the SPRSound dataset and obtained the result on the test dataset with a sensitivity of 0.5716, specificity of 0.7882, average score of 0.6799, harmonic score of 0.6626, and total score of 0.6756.

Keywords: respiratory sound classification, CNN, LSTM, upsampling, embedding, FCNN.

1. Introduction

Respiratory diseases, including chronic obstructive pulmonary disease (COPD), asthma, occupational lung diseases, and pulmonary hypertension, are one of the fatal causes of mortality in the world. According to the World Health Organization, more than 3 million deaths were caused by respiratory diseases in 2019, which is still rising [1]. These respiratory diseases have already become a significant problem in people's daily life.

The stethoscope, invented by René Laennec in the 1860s, has been used to auscultate over centuries. Auscultation of the respiratory system is proven to be non-invasive, safe, inexpensive, and easy to perform [2]. However, auscultation suffers from intrinsic limitations such as variability and subjectivity [3]. The examination must be conducted face-to-face, and it needs expert physicians that are highly experienced. According to Salvatore et al., hospital trainees and medical students misidentified half of the pulmonary sounds [4]. These limitations of auscultation need to be addressed by a system that can automate the auscultation process with high accuracy, high speed, and low cost.

With the rapid development of technology, deep learning has now been widely applied to medical fields. Massive, recorded data for respiratory sound are available, which makes data-driven deep learning for respiratory sound classifications more feasible. Many studies have already proposed various architectures of deep learning networks and machine learning algorithms to address this problem.

This paper proposed a novel machine-learning architecture that takes advantage of upsampling

DOI: 10.54254/2753-8818/28/20230401

Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM), and Fully Convolutional Neural Network (FCNN) to classify respiratory sounds. The upsampling CNN extracts and magnifies the feature from the input, and the output is fed into the LSTM which is responsible for recognizing and memorizing the long-term dependencies of the features. At the output, an FCNN combined with embedding layers will learn the relationship between the input signal and metadata (age, gender, position). The model will be trained and evaluated on the SPRSound dataset, which includes seven types of lung sounds (Normal (N), Rhonchi (R), Wheeze (W), Stridor (S), Coarse Crackle (CC), Fine Crackle (FC), Wheeze & Crackle (WC)).

2. Related work

There have been numerous studies of automating lung sound classification using machine learning. For example, Yoonjoo et al. explored the idea of using CNN to classify lung sounds into 4 categories, but the accuracy varies significantly for different sample groups [5]. Similar studies have been made in heart sound classifications by Gari et al. [6]. Even more, studies have used ML to classify any type of sound. Aditya et al. examine using CNN and Tensor Deep Stacking Network (TDSN) in sound classification [7]. Pablo et al. has studied the effectiveness of LSTM in classifying broadcast domain data [8]. A more recent study by Georgios et al. analyzed the effectiveness of a hybrid CNN-LSTM network with focal loss function in classifying lung sounds, and they achieved an accuracy of around 75% [9].

3. Methodology

3.1. Dataset

The dataset used for this paper is the SPRSound dataset. This dataset collects children's lung sounds whose ages range from 0-18, in four different positions of the body: left posterior (p1), left lateral (p2), right posterior (p3), and right lateral (p4). The data is recorded in '.wav' format with a sample rate of 8000. The annotations of the dataset are organized in '. json' format and provide labels at the event level and record level. There are 5 classes (Normal, CAS, DAS, CAS & DAS, or Poor Quality) at record level based on the presence/absence of continuous/discontinuous adventitious respiratory sounds, and 7 classes (Normal (N), Rhonchi (R), Wheeze (W), Stridor (S), Coarse Crackle (CC), Fine Crackle (FC), Wheeze & Crackle (WC)) at the event level. One audio file contains several segments of respiratory events. The start time and end time of the segments are specified in the file label, in milliseconds [10][11]. In this paper, only the dataset at the event level is used (Respiratory Sound Classification at Event Level). Only the training and testing datasets are available in the repository, so the training dataset is split into 70% and 30% for training and validation, respectively. The final performance of the model is evaluated using the testing dataset.

3.2. Data preprocessing

The audio sample was first sliced into different pieces according to the start time and end time in the label file as a single data. The numbers of each class are shown in Table 1 below. The audio samples of the breathing sound are converted from time-volume representation to 3 similar spectrogram representations: mel-spectrum, LFCC (linear-frequency cepstrum coefficients), and MFCC (Mel-frequency cepstrum coefficients). The difference between the 3 representations is the spacing of the frequency bandpass filters.

Table 1. The number of each class.

	Normal (N)	Rhonchi (R)	Wheeze (W)	Stridor (S)	Coarse Crackle (CC)	Fine Crackle (FC)	Wheeze & Crackle (WC)
Number	5159	39	452	15	49	912	30

DOI: 10.54254/2753-8818/28/20230401

3.3. Convolutional layer

CNN utilizes polling layers to condense input data, which filters out excess data that has little or no contribution to the overall learning. This also reduced the number of calculations in the downstream layer. By having multiple pooling layers, the core features are extracted, and the noise is stripped out. The opposite of the polling layers is the upsampling layer, which increases the input size. One method is duplicating the same value in each dimension.

3.4. RNN and LSTM layer

Recursive Neural Network (RNN) is suitable for dealing with serial input data [12]. The loopback of the information in the hidden layer emphasizes the relation between the inputs in the time domain [12]. LSTM utilizes the cell state to dynamically adjust the influence of previous input data [13]. The Bidirectional Long Short-Term Memory Network (Bi-LSTM) utilizes both a forward LSTM and a backward LSTM to capture potential relationships between the past input, current input, and future input. However, based on the assumptions made on the input data, the past input will have a more dominant effect on the current input than the future. The Bi-LSTM is not used to reduce computation. Thus, the effectiveness of the backward LSTM is not explored in this paper.

3.5. Embedding

The Hidden embedding layers are used to group inputs that have similar features. This grouping is done by adjusting the weight assigned to each input during the training process. The metadata is feed into the embedding layers to enhance the input. For example, the gender of the subject might affect the pitch and the volume of the recorded audio. There might be more features, which can be inferred from the gender of the subject, affecting the classification of the audio. Ideally, a higher dimension for the embedding layer could extract more features, but it requires a significant amount of training data to fine-tune the weights [14]. This paper uses 64-dimension embedding due to the limited training data available.

3.6. The proposed model

To compensate for the highly unbalanced dataset, a hybrid 2D upsampling convolutional LSTM with embedding layers network is used. The model receives a 3-dimensional normalized Mel-spectrogram concatenated with Mel-Frequency Cepstral Coefficients (MFCC) and Linear Frequency Cepstral Coefficient (LFCC) as input. These input data are classified into seven distinct classes: normal, rhonchi, wheeze, stridor, coarse crackle, fine crackle, and wheeze& crackle.

As shown in Figure 1 below, the whole model consists of three parts, the CNN part, the LSTM part, and the FCNN part. The CNN layers take the 3-dimensional normalized Mel-spectrogram together with MFCC and LFCC as input. There are 4 CNN layers in total. The ReLU activation function and batch normalization are applied after each CNN layer. Max-pooling layers emphasize the features captured from the CNN layers. Upsampling layers are used after the second and third CNN layers. Each convolution layer increases the number of channels at its output in the following order, 8, 16, 32, and 64. The result from the CNN part is a 3D array with a shape of 64x16x19.

Next, the result from the CNN part is reshaped and permuted into a 2D array with the shape of 304x64x1. These values are fed into a single LSTM layer with 512 hidden states. The goal of this LSTM layer is to recognize and memorize the long-term dependency and pattern from the input feature.

As mentioned in the previous part, the metadata (age, gender, position) is also an essential factor in classifying respiratory sounds. To learn the relation between the metadata and respiratory sounds, a fully connected neural network is added at the end of the model. In the FCNN part, the output from the LSTM is flattened into a 1-dimensional array and is passed into four fully connected layers. Each of the metadata (age, gender, position) is first passed into an embedding layer of size 64, and the outputs are concatenated with the result of the second fully connected layers. The Tanh activation function is used after the first and third fully connected layers to add non-linearity to the process. In the last fully connected layer, the number of output features is 7, corresponding to 7 target classes (normal, rhonchi, wheeze, stridor, coarse crackle, fine crackle, wheeze+crackle). The softmax is applied after the last layer

DOI: 10.54254/2753-8818/28/20230401

to retrieve the probability for each class. The cross-entropy error with class weight is used to compensate for the imbalanced dataset.

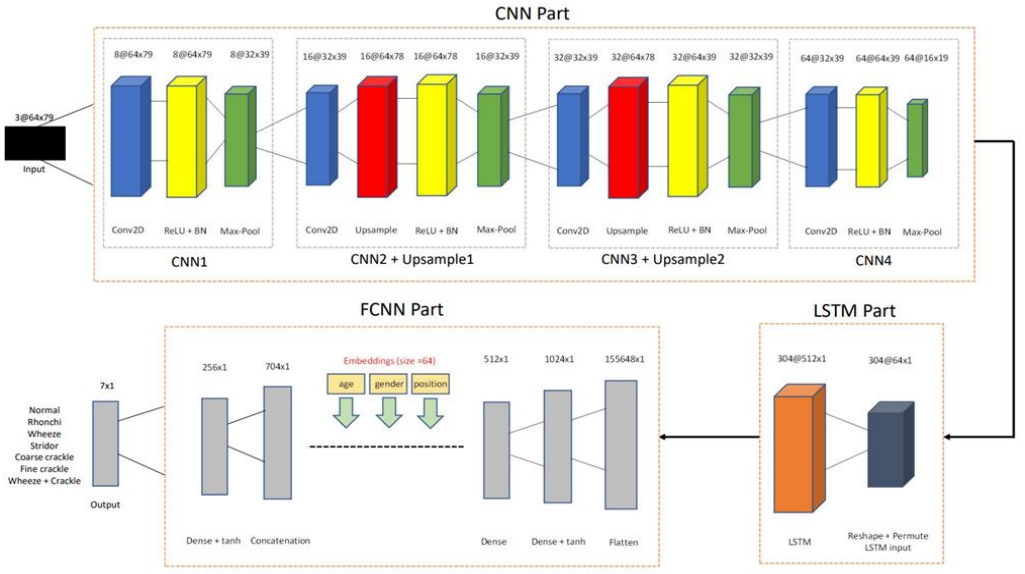

Figure 1. Structure of the proposed model.

4. Result and ablation study

This section analyzes the effect of different parts of the model by isolating each of them from the model. This section will be structured as follows: Section 4.1 indicates the data used for validation and testing; Section 4.2 specifies the metrics used to evaluate the model performance; Section 4.3 discusses the results of using only the CNN part of the model without upsampling layer, the results of using the CNN part of the model with upsampling layer, and the results of using upsampling CNN part with the LSTM part of the model (without the FCNN part).

4.1. Dataset

The same dataset (SPRSound dataset) is used to evaluate the performance of the model. As discussed in Section 3.1, 30% of the training dataset will be used for validation and the testing dataset will be used separately for testing. The performances on the validation and the testing dataset of different models are listed below.

4.2. Metrics

As discussed in Section 2, the dataset is highly unbalanced. The accuracy of major classes will significantly impact the overall accuracy, so additional metrics should be introduced to evaluate the overall performance of the dataset, which are sensitivity (Se), specificity (Sp), average score (AS), harmonic score (HS), and final score (TS). They are calculated as follows:

$$S_e = \frac{R_p + W_p + S_p + CC_p + FC_p + WC_p}{R_t + W_t + S_t + CC_t + FC_t + WC_t}$$

$$S_p = \frac{N_p}{N_t}$$

$$AS = \frac{S_e + S_p}{2}$$

$$HS = \frac{2\,S_e \cdot S_p}{S_e + S_p}$$

DOI: 10.54254/2753-8818/28/20230401

$$TS = \frac{S_e + S_p + AS + HS}{4} \quad (1)$$

Where Xp denotes the number of correct predictions of the X class, and Xt denotes the total number of samples in the X class (for example, Np is the correct number of predictions of the Normal (N) class, and Nt is the total number of Normal (N) class). These metrics consider the accuracy of both minority and majority classes, providing a better interpretation than overall accuracy in evaluating the performance of different models.

4.3. Result of different parts of models

In this part, the result of using different parts of the model will be compared and discussed. Multiple experiments were conducted to find the optimal hyperparameters for each different part of the models. The metric results mentioned in section 4.2 on the validation and the testing datasets are shown in Tables 2&3 below.

Table 2. Result of different models on the validation dataset.

	Se	Sp	AS	HS	TS
CNN	0.3623	**0.7719**	0.5671	0.4926	0.5298
Upsampling CNN	0.4262	0.6794	0.5528	0.5238	0.5383
Convlstm	**0.6408**	0.6533	0.6471	0.6467	0.6469
Convlstm + FCNN (full model)	0.5773	0.7584	**0.6679**	**0.6556**	**0.6638**

Table 3. Result of different models on the test dataset.

	Se	Sp	AS	HS	TS
CNN	0.4482	0.7604	0.6043	0.5640	0.5842
Upsampling CNN	0.5631	0.6291	0.5961	0.5943	0.5952
Convlstm	0.5149	0.7049	0.6099	0.5951	0.6025
Convlstm + FCNN (full model)	**0.5716**	**0.7882**	**0.6799**	**0.6626**	**0.6756**

Indicating by the Table 2 and the Table 3 above, the overall performance specified by the total score (TS) of the models is increasing in both the validation dataset and the testing dataset, after gradually adding each part. The CNN part of the model has already achieved high accuracy in predicting normal class, which is indicated by the high specificity. However, it has low accuracy in predicting minority classes (Rhonchi (R), Wheeze (W), Stridor (S), Coarse Crackle (CC), Fine Crackle (FC), Wheeze & Crackle (WC)), as indicated by the sensitivity. The harmonic score (HS) is enhanced when more components are added to the model, which suggests that the model is balancing the performance between the normal class and other classes. Finally, the total score (TS) and harmonic score (HS) show that the proposed model performs the best in comparison to other simpler models.

5. Conclusion

In this study, a hybrid upsampling CNN-LSTM with metadata embeddings model is proposed to classify the respiratory sounds into 7 categories. More specifically, respiratory sounds are transformed into mel-spectrogram images together with MFCC and LFCC and passed in an upsampling CNN to extract and emphasize the local features of the input. Then the features were fed into an LSTM network to memorize the long-term dependencies and patterns. Lastly, to learn the effect of metadata, the output of LSTM is

DOI: 10.54254/2753-8818/28/20230401

sent through an FCNN with metadata embedding layers. Additionally, the cross-entropy loss was used with class weight to compensate for the imbalance dataset.

Compared to simpler models, the proposed model showed better performance in both the validation and testing datasets. It demonstrated its ability to handle highly imbalanced datasets, especially in balancing the accuracy between the minority classes and the normal class. It obtained the highest total score in both the validation dataset and the test dataset (0.6638 and 0.6756, respectively).

For future studies, more experimentations are needed for uncovering potential improvement of the models. For example, using different pre-train models, such as ResNet, to replace the CNN upsampling layer when extracting the image features or explore the effectiveness of Bi-LSTM in audio classification.

References

[1] World Health Organization. (n.d.). Chronic obstructive pulmonary disease (COPD). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)

[2] Arts, L., Lim, E. H. T., van de Ven, P. M., Heunks, L., & Tuinman, P. R. (2020, April 30). The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary pathologies: A meta-analysis. Scientific reports. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7192898/#:~:text=Auscultation%20of%20the%20respiratory%20system,essential%20parts%20of%20clinical%20examination.

[3] Kim, Y., Hyon, Y., Lee, S., Woo, S.-D., Ha, T., & Chung, C. (2022, March 31). The Coming Era of a new auscultation system for analyzing respiratory sounds - BMC pulmonary medicine. BioMed Central. https://bmcpulmmed.biomedcentral.com/articles/10.1186/s12890-022-01896-1

[4] MANGIONE, S., & NIEMAN, L. Z. (n.d.). American Journal of Respiratory and Critical Care Medicine. https://www.atsjournals.org/doi/10.1164/ajrccm.159.4.9806083

[5] Kim, Y., Hyon, Y., Jung, S. S., Lee, S., Yoo, G., Chung, C., & Ha, T. (2021, August 25). Respiratory sound classification for Crackles, wheezes, and Rhonchi in the clinical field using Deep Learning. Nature News. https://www.nature.com/articles/s41598-021-96724-7

[6] G. D. Clifford et al., "Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in Cardiology Challenge 2016," 2016 Computing in Cardiology Conference (CinC), Vancouver, BC, Canada, 2016, pp. 609-612.

[7] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey and P. Tiwari. “Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network," in IEEE Access, vol. 7, pp. 7717-7727, 2019, doi: 10.1109/ACCESS.2018.2888882.

[8] Gimeno, P., Viñals, I., Ortega, A., Miguel, A., & Lleida, E. (2020, March 5). Multiclass audio segmentation based on recurrent neural networks for Broadcast Domain Data - EURASIP Journal on audio, speech, and music processing. SpringerOpen. https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-020-00172-6

[9] Petmezas, G.; Cheimariotis, G.-A.; Stefanopoulos, L.; Rocha, B.; Paiva, R.P.; Katsaggelos, A.K.; Maglaveras, N. Automated Lung Sound Classification Using a Hybrid CNN-LSTM Network and Focal Loss Function. Sensors 2022, 22, 1232. https://doi.org/10.3390/s22031232

[10] Q. Zhang, et al. “SPRSound: Open-Source SJTU Paediatric Respiratory Sound Database”, IEEE Transactions on Biomedical Circuits and Systems (TBioCAS), pp. 1-13, 2022, early access.

[11] Q. Zhang, et al. “Grand Challenge on Respiratory Sound Classification”, IEEE Biomedical Circuits and Systems Conference (BioCAS), 2022, pp. 1-5.

[12] Recurrent neural networks cheatsheet star. CS 230 - Recurrent Neural Networks Cheatsheet. (n.d.). https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

DOI: 10.54254/2753-8818/28/20230401

[13] Hochreiter, S., & Schmidhuber, J. (1997, November 15). Long short-term memory. MIT Press. https://doi.org/10.1162/neco.1997.9.8.1735
[14] Google. (n.d.). Embeddings | machine learning | google for developers. Google. https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture#:~:text=An%20embedding%20is%20a%20relatively,like%20sparse%20vectors%20representing%20words.