AUTOMATIC RESPIRATORY SOUND CLASSIFICATION VIA MULTI-BRANCH TEMPORAL CONVOLUTIONAL NETWORK

Ziping Zhao[1]*, Zhen Gong*[1]*, Mingyue Niu*[1]*, Jiali Ma*[1]*, Haishuai Wang*[2]*, Zixing Zhang*[3]*,Ya Li*[4]

[1]College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China
[2]Department of Computer Science and Engineering, Fairfield University, USA
[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[4]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

ABSTRACT

Automated classification of respiratory sounds has become an active research area in recent years. While recent studies have utilised deep learning methods to aid with respiratory sound classification, the performance is heavily influenced by the datasets available for respiratory sound classification tasks, which tend to be smaller and imbalanced. In this paper, we propose to explore the effectiveness of a multi-branch Temporal Convolutional Network (TCN) architecture integrated with Squeeze-and-Excitation Network (SEnet), a system denoted herein as MBTCNSE, for respiratory sound classification. To the best of the authors' knowledge, this is the first time that such a hybrid architecture has been employed for respiratory sounds classification. Experiments based on the ICBHI challenge respiratory sound dataset demonstrate the effectiveness of our method.

Index Terms— Respiratory sound classification, Temporal Convolutional Network (TCN), Squeeze-and-Excitation Network (SEnet)

1. INTRODUCTION

Respiratory diseases cause immense health, economic and social burdens. As the third leading cause of death worldwide, these represent a significant problem for public health systems [1]. Early diagnosis has been found to be crucial in limiting the spread of respiratory diseases, along with their adverse effects on the length and quality of life in most cases. Automatic respiratory sound classification thus been extensively explored by researchers; this is an approach that aims to expedite the process of machine-aided diagnosis for various respiratory conditions [1].

Early works focused on hand-crafted features and traditional machine learning [2, 3]. More recently, deep learning-based methods have been introduced for lung sound analysis [4, 5]. To train deep neural networks (DNNs), a time-frequency representation of the audio signal–such as Mel-spectrograms [4, 6], stacked MFCC features [7, 5] or an optimised S-transform spectrogram [8]–is employed. This 2D "image" is then fed into convolutional neural networks (CNNs) [7, 9], recurrent neural networks (RNNs) [5, 10], or hybrid CNN-RNNs [4] to facilitate the learning of robust high-dimensional representations.

Although recent works have achieved increasingly good performance in terms of the classification of respiratory sounds through the successful application of deep learning techniques, the majority of the literature [11, 12] has focused primarily on distinguishing healthy participants from abnormal patients. Furthermore, existing respiratory sound classification frameworks suffer from generalisation due to the limited number of training samples; their performance is limited to a small number of diseases/event types, and they have only been evaluated using a single dataset [13]. There is limited existing research on developing a lightweight model that is capable of learning from limited data while also having sufficient capacity to model temporal relationships across different time intervals.

As recent studies have shown, Temporal Convolutional Network (TCN) [14, 15] can be used to extract temporal information from the data provided at multiple resolutions, which can improve the system performance [16, 17]. In particular, TCN excel at capturing contextual information, meaning that they can achieve performance comparable or superior to that of Long Short-Term Memory (LSTM) across a diverse range of tasks and datasets [16]. Meanwhile, the representations produced by CNNs can be strengthened through the integration of a Squeeze-and-Excitation (SE) block; this is an architectural unit designed to improve a network's representational power by enabling it to perform dynamic channel-wise feature recalibration into the network, an approach that helps to capture the spatial correlations between features [18]. Furthermore, the multi-branch TCN has achieved better performance for monaural speech enhancement relative to single-branch TCN by incorporating one-dimensional causal dilated CNN and residual learning to expand receptive fields in or-

The present work is supported by the National Natural Science Foundation of China (No. 62071330), New Talent Project of Beijing University of Posts and Telecommunications (2021RC37) and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (No. 202200012).

der to capture temporal relationships at different granularities [19].

Motivated by the above observations, we propose to utilise multi-branch Temporal Convolution Network integrated with a Squeeze-and-Excitation Network (SEnet), namely MBTCNSE, which is a TCN-based model embedded with SEnet for respiratory sound classification. Our proposed approach exploits the advantages of both the RNN and CNN methods, including combining long-distance information with a flexible receptive field size, achieving real-time processing with variable length inputs, parallel computing convolution, and stable gradients in contrast to RNN methods [15]. Moreover, we employ the SE block proposed in [18] to enhance the network's representation ability. Finally, the majority voting results of these segments are taken as the result corresponding to the long-term audio. To the best of our knowledge, this is the first time that such a study has been conducted for respiratory sound classification.

Our key contributions can be summarised as follows. i) We have developed a multi-branch Temporal Convolutional Network architecture, integrated with a Squeeze-and-Excitation Network, for respiratory sound classification. ii) Extensive experiments on the ICBHI respiratory sound database demonstrate that the proposed model outperforms other state-of-the-art approaches.

2. PROPOSED METHOD

In order to automatically classify the respiratory sounds, we first divide the long-term audio into short-term segments with a fixed duration. The log mel spectrum feature is extracted from the corresponding short-term audio segment. The proposed multi-branch TCN model is then exploited to examine multi-scale information and predict the label of the audio segment. Finally, the majority voting results of these segments are taken as the result corresponding to the long-term audio. Fig. 1 illustrates the pipeline of our proposed method.

2.1. Log Mel Spectrum Feature from Audio Segment

As discussed above, the long-term audio is divided into a certain number of audio segments using a fixed-length window. We choose to adopt this processing strategy for two key reasons. First, we are able to conduct a detailed exploration of the differences between lung sounds from the short-term audio. Second, more samples can be obtained for use in training the proposed network model. It should be noted here that the label of the short-term audio segment is set to be the same as that of the corresponding long-term audio in our experiments.

2.2. Temporal Convolutional Network (TCN)

To better capture intrinsic time-frequency information from the spectrogram, a temporal convolutional network [15] is utilised to learn the temporal dynamics representation (Fig. 2). A TCN cell comprises three parts: *causal convolutions*, *dilated convolutions* and *residual connections*. In causal convolutions, information cannot be passed from the future to the past. Moreover, given that sequence modelling should be capable of looking "very far" into the past, dilated convolutions are employed to enable an exponentially large receptive field. In more detail, for a sequence input $\mathrm{x} \in \mathbb{R}^n$, the dilated convolution operation F on element s of the sequence is defined as follows:

$$F(s) = (x *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i}, \tag{1}$$

where d denotes the dilation factor, k respresents the filter size, and $s - d \cdot i$ accounts for the direction taken in the past. In Eq. (1), the use of a larger dilation enables an output at the top level to represent a wider range of inputs, which effectively expands the receptive field. This ensures that a filter exists that is capable of hitting every input within the effective history, while also allowing for an extremely large effective history using deep networks [15].

Moreover, in our work, we have added the channel-wise attention after the last $1 \times 1 - conv$ in the original TCN block using SENet [18].

2.3. Multi-branch TCN for Fusing Effective Information

Due to the assessment cues are being contained in different time ranges, multiple TCN branches are used to capture information in different receptive fields. More specifically, we input the log mel spectrum feature extracted from short-term audio into the multi-branch TCN, in combination with SENet, to capture the multi-scale information as shown in Eq. (2).

Similar to [13], the number of TCN branches is set to 3 in our work, since the multi-branch TCN model has been found to achieve optimal performance when the number of branches is set to 3.

$$\begin{cases} \mathbf{o}_1 = \mathrm{TCN1}\left(\mathbf{x}\right) \\ \mathbf{o}_2 = \mathrm{TCN2}\left(\mathbf{x}\right) \\ \mathbf{o}_3 = \mathrm{TCN3}\left(\mathbf{x}\right) \end{cases}, \tag{2}$$

where $\mathrm{TCN1}(\cdot)$, $\mathrm{TCN2}(\cdot)$ and $\mathrm{TCN3}(\cdot)$ denote three TCNs with different receptive field sizes.

Exploring the above process, on the one hand, we investigate the impact of the short-term audio on the target task across different temporal ranges. On the other hand, the information is enriched due to the construction of the three branches, which improves the recognition accuracy. The corresponding experimental results also verify the effectiveness of our proposed method.

Furthermore, in order to take advantage of the multi-branch TCN, we construct the outputs from the three TCN into a tensor with three channels and adopt a Conv2D layer to reduce the number of channels to 1 in the aggregate. Notably, in our experiments, we set the kernel of Conv2D to 1×1;

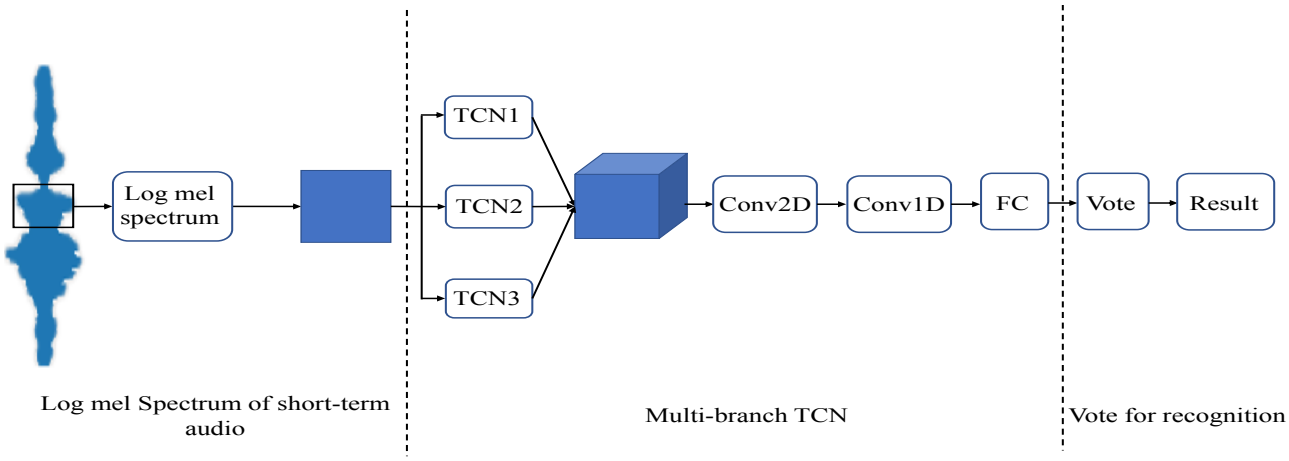

Fig. 1. Illustration of our proposed MBTCNSE model.

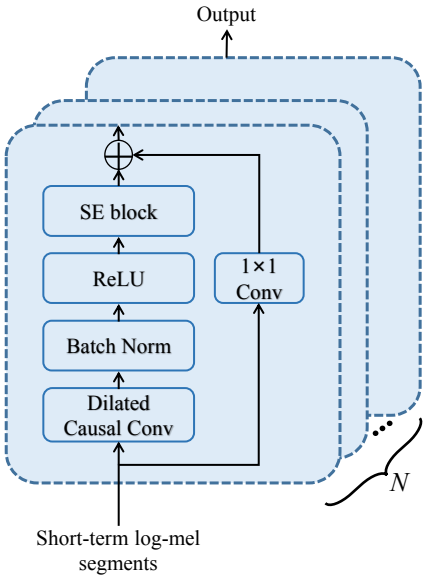

Fig. 2. Illustration of temporal convolutional network (TCN).

thus, this process is equivalent to a weighted summation of channels. Subsequently, a Conv1D layer and fully connected layer are adopted for recognition purposes.

3. EXPERIMENTS

In this section, we first briefly describe the database employed in the experiments, after which the experimental settings are given. Finally, the results and discussion are presented.

3.1. Databases and Evaluation Metrics

Experiments are conducted on the ICBHI scientific challenge respiratory sound dataset [20]. This is one of the largest publicly available respiratory datasets. The dataset comprises 920 audio recordings taken from a total of 128 patients with a combined total duration of 5.5 hours. Each breathing cycle in a recording is annotated by an expert belonging to one following four classes: *normal*, *crackle*, *wheeze*, *crackle & wheeze*. These cycles have various recording lengths (from 0.2s up to 16.2s), with the number of cycles being unbalanced, with 1864,886,506 and 3 642 cycles respectively for *crackle*, *wheeze*, *crackle & wheeze*, and *normal*. The full audio recordings have vary in length from 10 to 90s, and a wide spectrum of sampling frequencies are used ranging from 4 kHz to 44.1kHz.

We employ the standard benchmarks for evaluation in this paper: i.e., Accuracy (Acc.), Sensitivity (Sen.), Specificity (Spe.) and ICBHI score are used for evaluation in this paper [20]. We compare our performance using the official 80–20% train-test split [21].

3.2. Experimental Setup

Before feature extraction, we first resampled the recordings to 8kHz, and then divided each cycle into respiratory segments using a fixed-length window. We copied the cycles with durations of less than 3 sec and concatenated them to produce cycles with lengths of longer than 3 sec. The label of each segment is marked as the label of the original cycle.

We extracted the log mel-filter bank energy feature of the segments using 40 filters. In addition, we calculated 40 first-order and 40 second-order differences between these features. All features are then concatenated together, yielding a feature vector of dimension 120. For all feature extraction processes, we employed a window length of 0.02 sec and a step size of 0.01 sec. Before the features are fed into the model, z-score normalisation is utilised. Overlap is employed as our data augmentation method. In order to obtain more segments for model training, we used 75% overlap rate instead of 50%. In this way, we are able to explore the detailed differences in respiratory sound among subjects from the short-term log-mel

Table 1. Performance comparison between the proposed model and other models on the ICBHI respiratory sound database.

Methods	Acc.	Sen.	Spe.	Score
Bi-RestNet [21]	67.4	58.5	80.1	69.3
LSTM [10]	–	62.0	85.0	74.0
Parallel-pooling CNN [22]	71.2	57.7	83.2	70.5
CNN	70.2	56.5	81.3	68.9
TCN	71.4	62.3	85.5	73.9
MBTCN	72.1	63.2	85.8	74.5
CNNSE	71.3	57.4	82.5	70.0
MBTCNSE	**72.5**	**65.3**	**86.1**	**75.7**

feature segments.

Implementation Details: The number of residual blocks contained in TCN is denoted as n, the distance between each filter tap in a dilated convolution layer is d, the number of filters in each convolution layer is k, and the kernel sizes of the convolutional layers are used as hyper-parameters. Following model evaluation, these parameters were set to $n = 7$, $d = 2^n$ and $k = 64$, while the kernel size of the convolutional layers is set to 2. The proposed model was trained for 100 epochs using cross-entropy as the loss function and the Adam optimiser (mini-batch of size 64, and a learning rate of 10^{-4}).

3.3. Results and Discussion

This section presents the results of our experiments, with the aim of verifying the efficiency of our proposed MBTCNSE framework. We first performed an ablation analysis to elucidate the benefits of incorporating TCN, multi-branch, and the SE block into the final proposed model. The effectiveness of our hybrid framework is further highlighted through comparison with other key results obtained in the literature on the ICBHI dataset (see Table 1). The state-of-the-art models utilised for comparison purposes are listed in Table 1 and comprise three methods that have previously achieved good performance on the ICBHI dataset. The models that employ CNN are used alone, while the CNN model with SE block, the single-branch TCN, and our proposed model without SE block are also compared with our proposed approach.

From the results, it can be observed that the proposed approach outperforms previous works on the ICBHI dataset (cf. Table 1).

As for the Squeeze-and-Excitation Network introduced in this work, the performance of the model with the SE block removed can be observed to be lower than that achieved by the model with integrated SE block on the ICBHI dataset. Moreover, the CNN model with SE block performs better than the model with CNN alone, while the performance of our proposed model is superior to that of the MBTCN model. From these results, we can conclude that incorporating an SE block into a TCN model such as ours is an effective solution that is well-suited for respiratory sounds classification.

Furthermore, to determine the benefits of using multiple branches for respiratory sound classification, we additionally compared the performance of single-branch TCN with the multi-branch architecture. Table 1 shows the importance of considering dependencies across multiple temporal granularities.

In summary, the present results demonstrate that our proposed model achieves notable performance improvements on the ICBHI dataset compared to other existing methods, which in turn demonstrates the effectiveness of our proposed hybrid network.

4. CONCLUSIONS

In this paper, we introduced a novel deep learning framework for respiratory sound classification. We demonstrated how recent successes in multi-branch temporal convolution neural network-based temporal modelling can be incorporated to mitigate the challenges posed by data scarcity. Moreover, the SE block is adopted to highlight the useful channels. Experimental results on the ICBHI respiratory sound database illustrate the superiority of our method.

In the future, we will consider applying our proposed framework to analyse other one-dimensional signals in order to detect potential abnormalities in these signals.

5. REFERENCES

[1] B.M. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, et al., "A respiratory sound database for the development of automated classification," in *Proc. International Conference on Biomedical and Health Informatics (ICBHI)*, Thessaloniki, Greece, 2017, pp. 33–37.

[2] G. Chambres, P. Hanna, and M. Desainte-Catherine, "Automatic detection of patient with respiratory diseases using lung sound analysis," in *Proc. 2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, La Rochelle, France, 2018, pp. 1–6.

[3] N. Jakovljević and T. Lončar-Turukalo, "Hidden markov model based respiratory sound classification," in *Proc. International Conference on Biomedical and Health Informatics (ICBHI)*, Thessaloniki, Greece, 2017, pp. 39–43.

[4] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 3, pp. 535–544, March 2020.

[5] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *Proc. International Conference on Artificial Neural Networks (ICANN)*, Rhodes, Greece, 2018, pp. 208–217.

[6] R. Liu, S. Cai, K. Zhang, and N. Hu, "Detection of adventitious respiratory sounds based on convolutional neural network," in *Proc. 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Shanghai, China, 2019, pp. 298–303.

[7] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 1–9, Sep. 2017.

[8] H. Chen, X. Yuan, Z. Pei, M. Li, and J. Li, "Triple-classification of respiratory sounds using optimized s-transform and deep residual networks," *IEEE Access*, vol. 7, pp. 32845–32852, March 2019.

[9] D. Perna, "Convolutional neural networks learning from respiratory data," in *Proc. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Madrid, Spain, 2018, pp. 2109–2113.

[10] D. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *Proc. 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, Cordoba, Spain, 2019, pp. 50–55.

[11] L.D. Pham, H. Phan, R. Palaniappan, A. Mertins, and I. McLoughlin, "Cnn-moe based framework for classification of respiratory anomalies and lung disease detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 2938–2947, Aug. 2021.

[12] G. Altan, Y. Kutlu, and N. Allahverdi, "Deep learning on computerized analysis of chronic obstructive pulmonary disease," *IEEE journal of biomedical and health informatics*, vol. 24, no. 5, pp. 1344–1350, May 2019.

[13] T. Fernando, S. Sridharan, S. Denman, H. Ghaemmaghami, and C. Fookes, "Robust and interpretable temporal convolution network for event detection in lung sound recordings," *arXiv preprint arXiv:2106.15835*, 2021.

[14] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 156–165.

[15] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[16] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, and B.W. Schuller, "Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition," *Neural Networks*, vol. 141, pp. 52–60, March 2021.

[17] C. Li, B. Chen, Z. Zhao, N. Cummins, and B.W. Schuller, "Hierarchical attention-based temporal convolutional networks for eeg-based emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, 2021, pp. 1240–1244.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, USA, 2018, pp. 7132–7141.

[19] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "Monaural speech enhancement using a multi-branch temporal convolutional network," *arXiv preprint arXiv:1912.12023*, 2019.

[20] B. M. Rocha, D/ Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni, et al., "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, vol. 40, no. 3, pp. 035001, 2019.

[21] Y. Ma, X. Xu, Q. Yu, Y. Zhang, Y. Li, J. Zhao, and G. Wang, "Lungbrn: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm," in *Proc. 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Nara, Japan, 2019, pp. 1–4.

[22] F. Demir, A. M. Ismael, and A. Sengur, "Classification of lung sounds with cnn model using parallel pooling structure," *IEEE Access*, vol. 8, pp. 105376–105383, June 2020.