Intelligent Stethoscope using Full Self-Attention Mechanism for Abnormal Respiratory Sound Recognition

Changyi Wu[1,†], Dongmin Huang[1,†], Xiaoting Tao[2], Kun Qiao[2], Hongzhou Lu[2,*] and Wenjin Wang[1,*]

***Abstract*— Machine learning automates the recognition of abnormal respiratory sounds and pulmonary diseases for wireless stethoscopes. However, most learning-based methods have unbalanced performance between low sensitivity (SEN) and high specificity (SPE). Recently, the full self-attention mechanism-based Transformer made significant progress in various medical tasks, but its role in respiratory sound recognition still remains unknown. It can extract the contextual information from segments with arbitrary length in a signal, especially with long-range dependencies. This is typically suitable for mining the pattern of temporally-continuous pathological respiratory sounds, including stridor, wheezes, and rhonchi. Thus in this paper, we explore the feasibility of using full self-attention mechanism of Audio Spectrogram Transformer (AST) to improve the performance of respiratory sound recognition, where FNN, CNN and AST are benchmarked on the dataset of ICBHI 2017. In our proposed framework, the input samples are generated by a new respiratory cycle-based segmentation in order to preserve the consistency of input representation; a dual-input AST model is designed to enhance the robustness to disturbances by extracting the complementary information between the spectrograms and log Mel spectrograms. Extensive experiments show that AST outperforms other methods in the task of respiratory sound recognition. Moreover, the proposed respiratory cycle-based segmentation considerably improves SEN by almost 10%.**

***Index Terms*— Stethoscope, respiratory sound, Transformer, signal segmentation, ICBHI 2017.**

I. INTRODUCTION

Respiratory diseases are a common cause of death worldwide, killing approximately 3 million people every year [1]. Clinically, early diagnosis of respiratory diseases can support medical decisions and optimize treatment solutions to improve the patients' survival rate. Auscultation is a crucial tool for examination and diagnosis, where physicians measure the respiratory sounds to assess and track the condition of pulmonary system of the patient. However, its diagnosis is often subjective, leading to a high number of missed diagnoses and misdiagnosis [2]. Therefore, automatic respiratory sound recognition based on machine learning has attracted much attention in the research of stethoscope [3].

This work is supported by the National Key R&D Program of China (2022YFC2407800), General Program of National Natural Science Foundation of China (62271241), Guangdong Basic and Applied Basic Research Foundation (2023A1515012983), and Shenzhen Fundamental Research Program (JCYJ20220530112601003).

[1]Department of Biomedical Engineering, Southern University of Science and Technology, China.

[2]Department of Thoracic Surgery, The Third People's Hospital of Shenzhen, China.

[†]These authors contributed equally to this work.

[*]Corresponding author: Hongzhou Lu (luhongzhou@fudan.edu.cn), Wenjin Wang (wangwj3@sustech.edu.cn)

In earlier studies [3], most researchers used shallow machine learning methods combined with handcraft features to recognize abnormal respiratory sounds. The emergence of deep learning boosts the exploitation of end-to-end deep-net models for respiratory sound recognition [4]–[6]. Pham et al. [7] presented an Inception-based deep neural network for recognizing pulmonary diseases using spectrograms.

Though the newly proposed methods made clear progress in the respiratory sound recognition, they still face some issues. First, convolutional neural networks (CNN) needs deeper networks or larger convolutional kernels to learn the lengthy contextual information from respiratory sounds in a time interval. This increases the number of parameters in the model and easily leads to overfitting in the condition of limited training data [9]. Second, although recurrent neural networks (RNN) can capture contextual information using hidden states, it is vulnerable to gradient disappearance and gradient explosion during training [10]. Consequently, the model's convergence is hindered and difficult to update the weights efficiently using training data [11]. Recently, the Transformer based on full self-attention mechanism has demonstrated success in various medical tasks [9], [10], which shows the capability of extracting contextual information in time sequences with arbitrary length, including long-range dependencies. We expect it to learn the patterns of continuous pathological respiratory sounds well, such as stridor, wheeze, and rhonchi [3]. However, its validity in respiratory sound recognition remains to be explored.

This paper, therefore, investigates the feasibility of using full self-attention mechanism to improve the abnormal respiratory sound recognition. The Audio Spectrogram Transformer [10] developed on top of such mechanism is used in this paper to learn representations from the spectrograms of respiratory sounds. The benchmark involving three methods was conducted on the dataset of ICBHI 2017 [12], including Fully Connected Neural Networks (FNN), CNN, and AST. In addition to the model innovation, we proposed a new respiratory cycle based segmentation to generate input samples in order to improve the consistency of input representation against the earlier methods that use fixed-length windows to generate input segments [8]. We also proposed a dual-input model that integrates spectrograms and log Mel spectrograms to enhance the robustness to the interference using the complementary information in between. Spectrograms emphasize the global information in the frequency domain while log Mel spectrograms prioritize the local low-frequency components. Such integration can provide a more comprehensive representation against

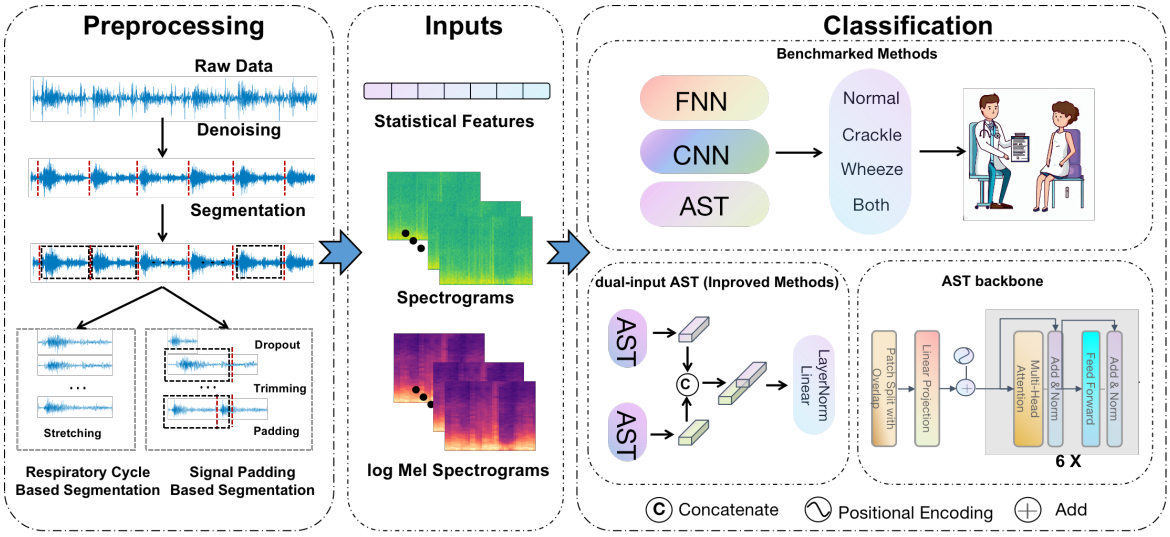

Fig. 1. Framework of the benchmarked methods and our proposed method. In the *preprocessing*, the respiratory cycle-based segmentation is proposed to stretch the respiratory cycle to a specific length and compare its effect with the sample padding method [8]. In the *classification*, FNN, CNN, and Audio Spectrogram Transformer (AST) were benchmarked based on ICBHI 2017. To improve the performance, we proposed a dual-input AST model.

the interference of noise. The experiments show that AST outperforms FNN and CNN, achieving 44.55% sensitivity (SEN) and 55.95% specificity (SPE). The dual-input AST model obtained the best performance with 42.91% SEN and 62.11% SPE. The main contributions of this paper are as follows:

1) The benchmark validates the full self-attention mechanism in the task of abnormal respiratory sound recognition. Moreover, the dual input AST model is proposed to improve the robustness of recognition.
2) The respiratory cycle based segmentation is proposed to improve the consistency of input samples, which shows a clear improvement on SEN in the recognition of abnormal respiratory sounds.

II. METHODOLOGY

The framework of benchmark is shown in Fig. 1. In the preprocessing, existing methods use sample padding [8] to segments the raw signal into short intervals with a fixed-length window, neglecting the characteristics of respiration. Particularly, it discards longer segments of the signal and replicates shorter segments by its ending samples. This may cause two issues: (i) the discarded part may contain important pathological information, such as wheezes in the latter half of respiratory cycle [14]; (ii) the self-replication may change the original sound patterns and cause ambiguity in the location of abnormal sound associated with certain diseases. For example, wheezes occurring in the expiration phase may indicate the presence of bronchial asthma and chronic obstructive pulmonary disease. When it appears in the inspiration phase, it may be related to the bronchial lumen narrowing or bronchospasm [3], [14]. Thus, self-replication may disrupt the phase information, leading to misdiagnosis. Therefore, we propose a new segmentation method that uses respiratory cycles as a unit to segment the respiratory sound signal consistently, ensuring that the input samples resemble the same amount of information for training and testing.

The benchmark includes three models, FNN, CNN, and AST, for abnormal respiratory sound recognition. The design of FNN and CNN refers to previous studies: (i) the statistical features, including zero-crossing rate, energy entropy, spectral centroid spread, spectral entropy, spectral flux, and MFCC [15], were extracted from the interval and FNN was applied to distinguish between different respiratory sounds [16]; (ii) CNN mined the task-related features from the (log Mel) spectrograms [2], [16], [17]. AST is developed based on the full self-attention mechanism [10], which features the property of efficiency in the extraction of contextual information of respiratory sounds. The single-input channel may not be sufficiently robust to disturbances [2]. Previous studies shew that integrating multiple features can mitigate the effects of noise [18]. Thus we are inspired to develop a dual-input AST model that integrates the complementary representations of spectrogram and log Mel spectrogram for a more comprehensive representation, which is expected to improve the performance of recognition.

A. Respiratory-cycle based Segmentation

To improve the consistency of input respiratory sound samples, we proposed a preprocessing method, called respiratory-cycle based segmentation, to segment the respiratory sound signal into short intervals in the unit of a respiratory cycle. Since respiratory cycles may have different time length, they are resampled to the same length. In this way, we mitigate the issue that an input sample may include incomplete or redundant phase information, i.e. all samples should resemble the same amount of information with the same phase for training and testing. Specifically, the preprocessing can be divided into three steps: (i) the raw signal is resampled at 4 kHz and filtered using 5-th order Butterworth high-pass filter with upper cut-off frequencies of 50 Hz; (ii) based on the start and end timestamps of each respiratory cycle, the raw signal is segmented into multiple intervals; (iii) the segmented intervals (or respiratory

TABLE I

BENCHMARK RESULTS OBTAINED BY COMBINATIONS OF DIFFERENT PREPROCESSING, MODELS, AND INPUT FEATURES. WE CONFIGURE THE STRETCHING TIME STRETCHING TO BE 5 S.

Preprocessing	Model	Features	ACC(%)	SEN(%)	SPE(%)	AS(%)	HS(%)
Respiratory cycle based segmentation	FNN	Statistical Features	45.49	18.15	**67.59**	42.87	28.61
	CNN	Spectrograms	46.85	43.69	49.40	46.55	46.37
		log Mel Spectrograms	47.13	40.19	52.74	46.46	45.61
	AST	Spectrograms	50.85	**44.55**	55.95	50.25	49.60
		log Mel Spectrograms	50.37	41.82	57.27	49.55	48.34
	AST-dual	Spectrograms+ log Mel spectrograms	**53.53**	42.91	62.11	**52.51**	**50.76**
Signal padding based segmentation [13]	FNN	Statistical Features	53.07	7.24	**84.30**	45.77	13.33
	CNN	Spectrograms	46.40	25.66	60.54	43.10	36.04
		log Mel Spectrograms	43.20	33.55	49.78	41.66	40.08
	AST	Spectrograms	51.73	33.55	64.13	48.84	44.05
		log Mel Spectrograms	**53.33**	33.55	66.82	50.18	44.67
	AST-dual	Spectrograms+ log Mel spectrograms	49.11	**40.40**	61.30	**50.85**	**48.70**

cycles) are stretched to a specific length; (iv) the magnitude normalization is carried out to the fixed-length intervals. In the experiment, the resampling length is set to 5 s. For benchmarking, the proposed segmentation is compared to the sample padding method [13] with the discarding length and truncation length 3 s and 5 s, respectively.

B. Deepl Models

1) FNN with the input of statistical features: The statistical features are extracted from respiratory sound signals to analyze the statistical attributes of respiratory sounds [15]. We employed FNN with two linear layers to process statistical features. The first linear layer has 102 input features and 32 output features. The second linear layer has 32 input features and 16 output features. The output layer is a linear layer that maps the 16 features of the feature layer to the number of classes.

2) CNN with the input of spectrograms and log Mel spectrograms: CNN extracts semantic features from 2D spectrograms of respiratory sounds through convolutional operations. According to [1], we employed the same network architecture using a single-channel spectrogram as the input and incorporating an augmented linear layer for the output.

3) AST with the input of (log Mel) spectrograms: AST can capture the contextual relationships between different time and frequency domains, leading to better recognition of respiratory sound patterns. In this paper, we utilized the pre-trained AST model [10], which is fully based on self-attention mechanisms for audio classification. The backbone is kept the same as the original paper, with modifications on the class number of output layer. Spectrograms and log Mel spectrograms were used as a single input to train and test AST.

C. Dual Input Architecture

Spectrograms and log Mel spectrograms are commonly used respiratory sound features. Log Mel spectrograms is more suitable for analyzing pitch variations in the detection of specific abnormalities like wheezes, while spectrograms provide higher frequency resolution that captures subtle frequency variations and detailed information in respiratory sounds. To exploit their complementary information, we employed two AST backbones with spectrogram and log Mel spectrogram features as separate input. The output two-dimensional features from each model are concatenated along the channel dimension. Subsequently, LayerNorm is applied to normalize the concatenated features, followed by a fully-connected layer for classification, as shown in the classification module in Fig. 1.

III. EXPRIMENTS AND DISCUSSION

A. Experimental setting

The ICBHI 2017 dataset [12] was used in this study, which consists of 920 recordings obtained from 126 patients using different stethoscopes on various body parts, amounting to a total duration of 5.5 hours. Each respiratory cycle in each audio record was labeled as normal, crackles, wheezes and both (wheezes and crackles). The database consists of a total of 6898 respiratory cycles, and the numbers of respiratory cycles for each type are 3642, 1864, 886, and 506.

The dataset was split as 60%/40% for training and test set in a subject-wise way according to the official division[1]. It should be noted that the data from one subject only appear either in the training set or test set. To evaluate the performance, multiple metrics were used, including accuracy (ACC), sensitivity (SEN), specificity (SPE), average score (AS), and harmonic score (HS), according to [12].

All methods were trained using the AdamW optimizer with an epoch of 300 and a learning rate of $10e^{-5}$. To address the issue of class imbalance in the dataset, the focal loss [19] was introduced that assigns lower weights to correctly classified samples and higher weights to misclassified samples.

B. Results

The benchmarked result is reported in Table I. It shows that AST with the input of spectrograms has 44.55% SEN and 55.95% SPE. It outperforms FNN and CNN, showing the highest score in the composite indexes of AS and HS. The improved performance may be attributable to the fact that AST uses the self-attention information of each region in the

[1]https://bhichallenge.med.auth.gr

spectrogram to mine the contextual information for updating the weights, which is unavailable in other models. It shows the feasibility of using the full self-attention mechanism to improve the abnormal respiratory sound recognition. Moreover, we vary the length of interval from 3 s to 6 s to explore the method stability, as shown in Fig. 2. It shows that AST-based methods achieved better performance in all settings, demonstrating the consistency of AST model in different settings.

Compared to the benchmarked methods, our proposed dual-input model, taking spectrograms and log Mel spectrograms as joint input, yields better performance with 42.91% SEN and 62.11% SPE. In comparison to the single input of (log Mel) spectrograms, it achieved the best average score among all models. The utilization of dual inputs enables the model to fuse complementary information from different feature representations, further enhancing robustness against disturbances.

Table I reported the comparison between our proposed preprocessing methods and [8]. As can be seen, the respiratory cycle-based segmentation can further improve the sensitivity of the employed models. This is due to the fact that segmentation based on signal padding involves trimming and replication the signal, thus disrupting the coherence of respiratory patterns. Our approach uses a respiratory cycle as a unit to generate input segments, which preserves information in a complete respiratory cycle, and most importantly, improves the consistency between different input samples.

In summary, the dual-input AST achieved the best performance among the benchmarked models. It demonstrates the effectiveness of fully self-attention mechanism in respiratory sound recognition. Furthermore, the dual-input model achieves the best performance with 42.91% SEN and 62.11% SPE by enhancing the consistency of representation. In the preprocessing, the proposed respiratory cycle based segmentation has an improvement of almost 10% in SEN.

IV. CONCLUSIONS

In this paper, we verified the feasibility of using full self-attention mechanism for abnormal respiration sound recognition. The AST model outperformed the widely used FNN and CNN models, and our proposed dual-input AST achieved the best performance. The proposed respiratory cycle-based segmentation can further improve the model's sensitivity to abnormal respiration sounds. These findings highlight the potential of using the full self-attention mechanism for stethoscopes to automate the diagnosis of respiratory disease, supporting and optimizing medical decisions in digital healthcare. In the future, the integration of larger and more diverse datasets can further enhance the model's applicability and robustness.

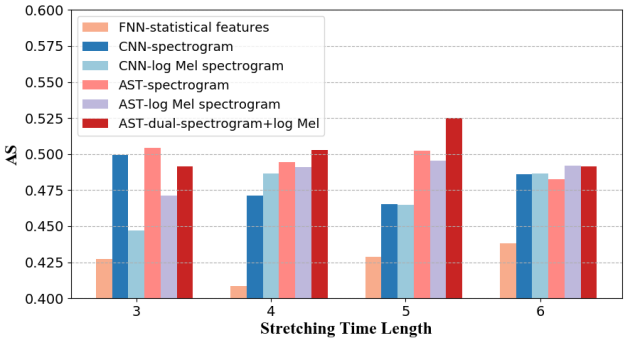

Fig. 2. Comparison of the average scores (AS) of different models when signals are stretched to different time lengths.

REFERENCES

[1] S. B. Shuvo *et al.*, "A lightweight cnn model for detecting respiratory diseases from lung auscultation sounds using emd-cwt-based hybrid scalogram," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2595–2603, 2020.

[2] B. M. Rocha *et al.*, "Automatic classification of adventitious respiratory sounds: A (un) solved problem?" *Sensors*, vol. 21, no. 1, p. 57, 2020.

[3] R. X. A. Pramono *et al.*, "Automatic adventitious respiratory sound analysis: A systematic review," *PloS one*, vol. 12, no. 5, p. e0177926, 2017.

[4] G. Petmezas *et al.*, "Automated lung sound classification using a hybrid cnn-lstm network and focal loss function," *Sensors*, vol. 22, no. 3, p. 1232, 2022.

[5] T. Nguyen and F. Pernkopf, "Lung sound classification using co-tuning and stochastic normalization," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 9, pp. 2872–2882, 2022.

[6] D. Huang *et al.*, "A multi-center clinical trial for wireless stethoscope-based diagnosis and prognosis of children community-acquired pneumonia," *IEEE Transactions on Biomedical Engineering*, 2023.

[7] L. Pham *et al.*, "Inception-based network and multi-spectrogram ensemble applied to predict respiratory anomalies and lung diseases," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 253–256.

[8] T. Nguyen and F. Pernkopf, "Lung sound classification using snapshot ensemble of convolutional neural networks," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*. IEEE, 2020, pp. 760–763.

[9] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] Y. Gong *et al.*, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[11] A. Rehmer and A. Kroll, "On the vanishing and exploding gradient problem in gated recurrent units," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1243–1248, 2020.

[12] B. M. Rocha *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, vol. 40, no. 3, p. 035001, 2019.

[13] T. Nguyen and F. Pernkopf, "Lung sound classification using snapshot ensemble of convolutional neural networks," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 760–763.

[14] S. İçer and Ş. Gengeç, "Classification and analysis of non-stationary characteristics of crackle and rhonchus lung adventitious sounds," *Digital Signal Processing*, vol. 28, pp. 18–27, 2014.

[15] G. Sharma *et al.*, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[16] W. Song *et al.*, "Contrastive embeddind learning method for respiratory sound classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2021, pp. 1275–1279.

[17] D. Huang *et al.*, "A contrastive embedding-based domain adaptation method for lung sound recognition in children community-acquired pneumonia," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

[18] T. Fernandes *et al.*, "Classification of adventitious respiratory sound events: A stratified analysis," in *IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE, 2022, pp. 01–05.

[19] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.