Interspeech 2024
1-5 September 2024, Kos, Greece

LungAdapter: Efficient Adapting Audio Spectrogram Transformer for Lung Sound Classification

Li Xiao[1], *Lucheng Fang*[2,†], *Yuhong Yang*[1], *Weiping Tu*[1,⋆]

[1]National Engineering Research Center for Multimedia Software,
School of Computer Science, Wuhan University
[2]Sleep Medicine Centre, Zhongnan Hospital of Wuhan University

tuweiping@whu.edu.cn

Abstract

Recently, fine-tuning the pre-trained large-scale Transformer models in lung sound classification tasks has yielded remarkable outcomes. However, the predominant method for fine-tuning is still full fine-tuning, which entails updating all parameters of large-scale models during training. Given the recent advancements in large-scale models, this approach requires significant computational resources and time. To tackle this issue, we introduce an efficient fine-tuning approach based on Adapter tuning, namely LungAdapter. This method can incorporate trainable blocks into a pre-trained audio Transformer model, allowing extraction of crucial information on lung sound classification from the model, while preserving the frozen parameters of large-scale pre-trained models. Experiments have shown that our method achieves performance comparable to or even superior to full fine-tuning while optimizing only 2.83% of the parameters.

Index Terms: Lung Sound Classification, Audio Spectrogram Transformer, Adapter, Fine-tune

1. Introduction

The incidence of chronic respiratory diseases (CRD), mainly chronic obstructive pulmonary disease (COPD), bronchitis and asthma, has gradually increased in recent years among children and adults [1]. Lung sounds contain abundant information about lung conditions and can be used to assess and diagnose chronic respiratory diseases [2, 3, 4]. The accurate early recognition of lung sounds is greatly significant for diagnosing lung diseases and assessing lung conditions. The advent of electronic stethoscopes has generated interest in contactless medical care and opened up new avenues for automated lung sound examinations, exemplified by the ICBHI dataset [5], which comprises lung sound recordings captured with electronic stethoscopes. Therefore, this technological advancement has not only facilitated the development of innovative diagnostic tools but also set the stage for applying deep learning techniques in analyzing lung sounds.

Deep learning-based methods for lung sound recognition, categorized by their neural network training strategies, are primarily divided into two types: training from scratch and fine-tuning. Training from scratch involves developing a model without pre-existing weights, only using ICBHI dataset [6, 7, 8, 9]. For example, LungRN+NL [6] explores data augmentation and ResNet [10] architecture to address the data class imbalance. In follow-up work, LungAttn [8] combines the attention mechanism and ResNet block to improve the classification accuracy of lung sounds. However, the small number of samples in the medical audio dataset leads to limited performance of methods trained from scratch. Conversely, the fine-tuning method leverages pre-trained models, which are fully fine-tuned on lung sound data to achieve higher performance. RespireNet [11] uses a ResNet model pre-trained on ImageNet [12], with a device-specific fine-tuning strategy. With the wide use of Transformer in various tasks, Bae *et al.* [13] use AST [14] with patch-mix augmentation and contrastive loss, which achieves significant advancements in lung sound classification tasks. Kim *et al.* [15] proposes a stethoscope-guided supervised contrastive learning approach, fine-tuning on the AST model and reducing shift in data distribution.

However, there are two primary limits to a full fine-tuning strategy. Firstly, pre-trained models have learned robust feature extraction ability through large-scale pre-training on high-quality data, and fine-tuning could lead to catastrophic forgetting. Secondly, as the size of these base models increases, so do the training costs. To address the limitations mentioned above, the parameter-efficient fine-tuning (PEFT) technique, such as Adapter, has been extensively studied in the fields of natural language processing (NLP) [16, 17, 18] and computer vision (CV) [19, 20, 21, 22] tasks. However, these methods have seen limited exploration in the field of lung sound recognition. The efficacy of the PEFT technique has inspired us to apply it to efficiently adapt a pre-trained audio Transformer for lung sound recognition tasks.

In this paper, we propose a method, called LungAdapter, to adapt tuning pre-trained AST models for lung sound recognition tasks efficiently. LungAdapter employs two distinct types of adapters: the Attention-Adapter and the Res-Adapter. The Attention-Adapter is positioned after the Multi-Head Self-Attention (MHSA) layer within the Transformer block. At the same time, the Res-Adapter operates in parallel to the MLP layer in a Transformer block. The methodology employed by adapters in [22] processes upstream features from multiple cognitive perspectives to enhance performance on downstream tasks. As a result, this has inspired us to utilize multiple convolutional filters in the Attention-Adapter to enhance the multi-scale information of the lung sound signal. Due to their careful design, our adapters can be seamlessly integrated into the pre-trained network. The experimental results demonstrate that our method achieves compelling transfer learning performance while maintaining a smaller parameter size.

The key contributions of this paper are summarized as follows:

• We introduce the LungAdapter method, which effectively transfers the pre-trained audio Transformer model to lung sound recognition tasks. This approach drastically reduces the training parameters of the model.

• Our multiple convolutional filters scheme in the Lun-

[†] Equal contribution. [⋆] Corresponding author.

10.21437/Interspeech.2024-106

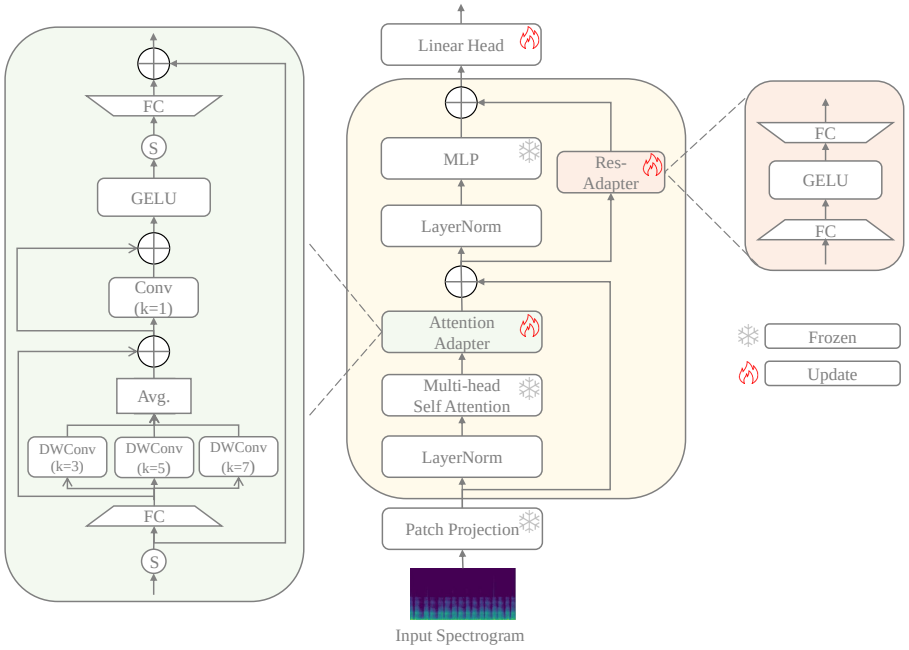

Figure 1: *Pipeline of our LungAdapter framework.* ***Middle****: Our framework consists of a frozen Audio Spectrogram Transformer (AST) augmented with trainable latent adapters inserted into each MHSA layer and MLP network. We use Attention-Adapter and Res-Adapter, which allow us to transfer information from audio/image to lung sound tokens.* ***Left****: Each Attention-Adapter consists of two fully connected (FC) layers and four convolutional filter layers with an intermediate activation layer.* ***Right****: The Res-Adapter includes two FC layers with an intermediate activation layer. We add multiple skip-connections in Attention-Adapter to minimize feature losses during convolution.*

gAdapter method captures multi-scale information of lung sound, ensuring better parameter transferring abilities.

• Experiments show that our method achieves highly competitive transfer learning performance while maintaining a relatively low level of trainable parameters.

2. Methodology

In this section, we first introduce the AST model applied in the ICBHI dataset [13] in Section 2.1. The pipeline of LungAdapter is described in Section 2.2. Then, we describe the proposed **Attention-Adapter** and **Res-Adapter** in detail, which are the key components of our model.

2.1. Audio Spectrogram Transformer in ICBHI

The AST has also been used in studies of lung sound classification, achieving state-of-the-art performance using fine-tuning approach [13]. Therefore, our method adapts the work in [13] as a backbone architecture.

Given the input feature map $\mathbf{X} \in \mathbb{R}^{T \times F \times C}$ as input, where T, F and C represent the time, frequency, and the number of channels in the spectrogram, respectively. To handle a 2D audio spectrogram, AST divides the feature map into a set of 16 $\times$ 16 spectrogram patches with an overlap of 6 in both time and frequency $[\mathbf{x}_p^1; \mathbf{x}_p^2; ...; \mathbf{x}_p^N]$, where $\mathbf{x}_p^i$ is the i-th patch of the input spectrogram. Therefore, there are $N = 12[(100t - 16)/10]$ patches for t seconds of audio, which are used as the input of the transformer model. Similar to [23], AST introduces an additional learnable positional embedding $\mathbf{E}_{pos}$ to retain positional information. This process can be formulated as:

$$\mathbf{X}_0 = \left[\mathbf{x}_{cls}; \mathbf{x}_p^1\mathbf{W}; \mathbf{x}_p^2\mathbf{W}; \ldots; \mathbf{x}_p^N\mathbf{W}\right] + \mathbf{E}_{pos} \quad (1)$$

where $\mathbf{W}$ is a learnable linear projection parameter, Then, the resulting sequence of embeddings serves as the input to the Transformer encoder.

The encoder in a Transformer consists of L attention blocks. Each block contains a multi-head self-attention (MHSA) [24] layer and an MLP network. For each layer, layer normalization (LN) [25] and residual connections [10] are employed. we can obtain the interacted spectrogram patch:

$$\mathbf{X}'_{l-1} = \mathbf{X}_{l-1} + \mathrm{MSA}(\mathrm{LN}(\mathbf{X}_{l-1})) \quad (2)$$

$$\mathbf{X}_l = \mathbf{X}'_{l-1} + \mathrm{MLP}\left(\mathrm{LN}(\mathbf{X}'_{l-1})\right) \quad (3)$$

where $l \in [1, ..., L]$ is the index of attention blocks. In order to perform classification tasks, the MHSA output is given to the feed-forward layer followed by a layer normalization and fully connected (FC) layers to classify the input. In this way, the output prediction $\mathbf{y}$ can be computed by:

$$\mathbf{y} = \mathrm{FC}\left(\mathrm{LN}\left(\mathbf{X}_L\right)\right) \quad (4)$$

In order to solve the scarcity of medical data, Bae *et al.* [13] based AST architecture also proposed patch-mix augmentation and contrastive loss.

2.2. LungAdapter

2.2.1. Attention-Adapter

There is a diverse range of feature domain information across various scales in the field of audio processing, including medi-

cal audio tasks. This has inspired us to utilize a multi-scale approach in the adapter to analyze the complex information of the signal. Therefore, we introduce the Attention-Adapter to capture multi-scale information [22] of lung sounds from the pre-trained model. As shown in the left of Figure 1, the Attention-Adapter employs a multi-view architecture consisting of two fully connected (FC) layers and four convolution layers with an intermediate activation layer. This design allows for the processing of upstream features from multiple cognitive perspectives, improving performance on lung sound classification tasks.

We introduce multiple convolutional filters to Attention-Adapter to increase the multi-scale information. Specifically, the output of the MHSA block goes through three depth-wise convolutional (DWConv) filters with kernel sizes of 3, 5, and 7 after the down projection. The output from these filters is averaged, and features are subsequently aggregated using a convolution with a kernel size of 1. Non-linearity is achieved through the Gaussian Error Linear Unit (GeLU), ultimately restoring the feature dimension to its original shape. The scale factor s regulates the distribution of features in the network. This process can be formulated as:

$$\mathbf{h} = \mathbf{W}_{\text{down}}\left(s_1 \cdot \mathbf{X}\right) \tag{5}$$

$$\mathbf{h}' = \mathbf{h} + \frac{1}{3} \cdot \sum_{k=3,5,7} \mathbf{DWConv}_k(\mathbf{h}) \tag{6}$$

$$\mathbf{y} = \mathbf{W}_{\text{up}}[s_2 \cdot \text{GeLU}(\mathbf{Conv}(\mathbf{h}') + \mathbf{h}')] + s_1 \cdot \mathbf{X} \tag{7}$$

Here, $\mathbf{X}$ denotes an input tensor, $\mathbf{y}$ is the output of Attention-Adapter. The $\mathbf{W}_{\text{up}}$ and $\mathbf{W}_{\text{down}}$ are the learnable matrix, and s is scale factor. The $\mathbf{h}$ and $\mathbf{h}'$ are hidden space features. The $\mathbf{DWConv}_k$ is depthwise convolution.

2.2.2. Res-Adapter

To extract lung sound information from the deep architecture of Transformer blocks, we utilize the Res-Adapter module for deep modeling, representing the learnable structures that are connected in parallel to the existing operations. With this unified formulation, the design of the adapter structure can learn class-related of lung sounds from the base model. As shown in right of Figure 1, the Res-Adapter includes two fully connected (FC) layers with an intermediate activation layer. The first FC layer maps the input to a lower-dimensional space, while the second FC layer maps it back to the original dimension. Formally, for an input feature matrix $\mathbf{X}$, the Res-Adapter could be written as:

$$\text{Res-Adapter}\left(\mathbf{X}\right) = \mathbf{W}_{\text{up}}\left(\text{GeLU}\left(\mathbf{W}_{\text{down}}\left(\mathbf{X}\right)\right)\right) \tag{8}$$

During training, all layers of the attention block are fixed, and only the adapters are updated.

3. Experiments

3.1. Experimental Settings

3.1.1. Dataset

We evaluate the performance of LungAdapter on the respiratory anomaly classification task proposed in the ICBHI challenge [5] organized at Int. Conf. on Biomedical Health Informatics. This is further divided into two subtasks: (1) classify a breathing cycle into one of the four classes– normal, crackle, wheeze, and both (crackle and wheeze), and (2) classify a breathing cycle into normal or anomalous class, where anomalous = crackle, wheeze, both. The ICBHI dataset comprises 6,898 respiratory cycles, which have a duration of approximately 5.5 hours and are officially split into a train set (60%) and a test set (40%). The train set includes a total of 539 recordings taken from 79 patients, that consist of 1,215 cycles of crackles, 501 cycles of wheezes, 363 cycles of both crackles and wheezes, and 2,063 cycles of normal breathing. Similarly, the test set includes 381 recordings taken from 49 patients, with a total of 649 cycles of crackles, 385 cycles of wheezes, 143 cycles of both crackles and wheezes, and 1,579 cycles of normal breathing. Following the evaluation metrics of the ICBHI 2017 challenge, we evaluate the classification performance by **Score**, the average of *Specificity* (S_p) and *Sensitivity* (S_e).

3.1.2. Implementation details

Following Bae *et al.* [13], we resampled all lung recordings to 16 kHz and set each duration to 8 seconds. Then, we extract log Mel spectrograms, with a frequency dimension of 128, using a 25 ms Hamming Window and a hop length of 10 ms. This gives us a resultant input size of 128 × 100 t for t seconds of audio. We also applied the standard normalization on the spectrograms with the mean and standard deviation of –4.27 and 4.57, respectively.

In this work, we use PyTorch toolkit [31] to conduct all experiments on NVIDIA RTX 4090 GPU. For most experiments, we adapt AST pre-trained from Bae *et al.* [13] on ImageNet [12] and Audioset [32] as the backbone model. For our LungAdapter model, The experimental setup mostly follows the [13]. We employed the Adam optimizer [33] with a weighted cross-entropy loss function, utilizing a learning rate of 5e-4, cosine scheduling, and a batch size of 8. We fine-tune the pre-trained AST model from [13] for only 50 epochs. The hidden space dimension of the adapter in Attention-Adapter is 16, and the bottleneck ratio of Res-Adapter is 0.15. It is worth noting that we present the results of our experiments over five random runs.

3.1.3. Baseline Methods

We compare LungAdapter with multiple recent methods. Baseline models can be grouped into methods without or with pre-trained models:

(1) Without pre-trained models: CNN-MoE [26], Ren *et al.* [9] and Chang *et al.* [27].

(2) With pre-trained models: Wang *et al.* [28], Nguyen *et al.* [29], Mouummad *et al.* [30], Bae *et al.* [13] and Kim *et al.* [15].

3.2. Results

3.2.1. Overall ICBHI Dataset Results

As presented in Table 1, our LungAdapter achieves competitive results (62.40% v.s. 62.37%), while we only need 2.48 M trainable parameters, which is less than one-thirty-five of the full fine-tune on AST. This confirms the effectiveness of our proposed pre-trained model adaptation strategy. For the 2-class classification in the ICBHI benchmark, we retrain the LungAdapter model instead of employing the pre-trained classifier from the 4-class task (e.g. Bae *et al.* [13] and Kim *et al.*[15]) due to its low memory usage. Among all experimental methods, LungAdapter performs the best and achieves 69.53% compared to the state-of-the-art method, but it only requires 2.48 M trainable parameters in the training process. Both versions of the lung sound classification task showcase the efficiency of our

Table 1: *Comparison of LungAdapter with state-of-the-art methods on ICBHI dataset. We compared it to previous works (the results from [15]) that use the official split of the ICBHI dataset. $*$ denotes the previous state-of-the-art Score.* ***Best*** *and* <u>*second best*</u> *results.*

	Method	Architecture	Pretrain	Venue	$S_p(\%)$	$S_e(\%)$	Score (%)	Trainable/Total Params (M)
4-class eval.	CNN-MoE [26]	C-DNN	-	*JBHI'21*	72.40	21.50	47.00	-
	Ren *et al.* [9]	CNN8-Pt	-	*ICASSP'22*	72.96	27.78	50.37	-
	Chang *et al.* [27]	CNN8-dilated	-	*INTERSPEECH'22*	69.92	35.85	52.89	-
	Chang *et al.* [27]	ResNet-dilated	-	*INTERSPEECH'22*	50.22	**51.83**	51.02	-
	Wang *et al.* [28](Splice)	ResNeSt	IN	*ICASSP'22*	70.40	40.20	55.30	-
	Nguyen *et al.* [29](Cotuning)	ResNet50	IN	*TBME'22*	79.34	37.24	58.29	-
	Mouummad *et al.* [30](SCL)	CNN6	AS	*arXiv'22*	75.95	39.15	57.55	-
	Bae *et al.* [13](Fine-tuning)	AST	IN+AS	*INTERSPEECH'23*	77.14	41.97	59.55	87.53/87.53
	Bae *et al.* [13] (Patch-Mix CL)	AST	IN+AS	*INTERSPEECH'23*	**81.66**	43.07	<u>62.37</u>*	87.53/87.53
	Kim *et al.* [15] (DAT)	AST	IN+AS	*ICASSP'24*	77.11	42.50	59.81	87.53/87.53
	Kim *et al.* [15] (SG-SCL)	AST	IN+AS	*ICASSP'24*	79.87	43.55	61.71	87.93/87.93
	LungAdapter(Ours)	AST	IN+AS	*INTERSPEECH'24*	<u>80.43</u>$_{\pm 3.11}$	<u>44.37</u>$_{\pm 2.46}$	**62.40**$_{\pm 1.11}$	**2.48**/90.01
2-class eval.	CNN-MoE [26]	C-DNN	-	JBHI'21	72.40	37.50	54.19	-
	Nguyen *et al.* [29] (Cotuning)	ResNet50	IN	*TBME'22*	79.34	50.14	64.74	-
	Bae *et al.* [13] (Fine-tuning)	AST	IN+AS	*INTERSPEECH'23*	77.14	56.40	66.77	87.53/87.53
	Bae *et al.* [13] (Patch-Mix CL)	AST	IN+AS	*INTERSPEECH'23*	**81.66**	55.77	68.71	87.53/87.53
	Kim *et al.* [15] (DAT)	AST	IN+AS	*ICASSP'24*	77.11	56.98	67.04	87.53/87.53
	Kim *et al.* [15] (SG-SCL)	AST	IN+AS	*ICASSP'24*	<u>79.87</u>	57.97	<u>68.93</u>*	87.93/87.93
	LungAdapter(Ours)	AST	IN+AS	*INTERSPEECH'24*	69.41$_{\pm 3.85}$	**69.65**$_{\pm 3.03}$	**69.53**$_{\pm 0.48}$	**2.48**/90.01

Table 2: *LungAdapter Design. We investigate different design choices of our method on the lung sound classification task.* ***1*** *and* ***2*** *are Attention-Adapter and Res-Adapter, respectively.*

Method	1	2	$S_p(\%)$	$S_e(\%)$	Score (%)	Trainable/Total Params (M)
Ours	✓	✓	80.43$_{\pm 3.11}$	44.37$_{\pm 2.46}$	**62.40**$_{\pm 1.11}$	**2.48**/90.01
	✓	✗	77.72$_{\pm 4.87}$	**45.05**$_{\pm 3.38}$	61.39$_{\pm 1.75}$	0.34/87.87
	✗	✓	77.55$_{\pm 2.38}$	44.55$_{\pm 2.25}$	61.05$_{\pm 0.51}$	2.14/89.67
	✗	✗	**81.66**$_{\pm 3.83}$	43.07$_{\pm 2.80}$	62.37$_{\pm 0.61}$	87.53/87.53

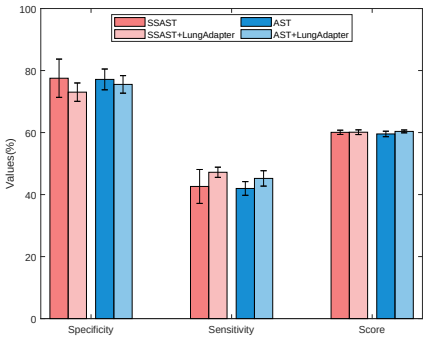

Figure 2: *Comparison of LungAdapter with different Transformer based methods on ICBHI dataset. The error bar represents the standard deviation.*

method in terms of parameter utilization and achieve comparable or superior performance compared to other methods.

3.2.2. Ablation studies

We verify the effectiveness of the two modules in LungAdapter: Attention-Adapter and Res-Adapter. As shown in Table 2, compared to LungAdapter with one of the mentioned modules, including two adapters leads to substantial performance enhancements of our model on lung sound classification tasks, demonstrating the effectiveness of these modules. Although any adapter can significantly reduce learnable parameters, we argue that distinct adapters can also mutually enhance each other's performance. For instance, removing any adapter module individually results in a significant decrease in the performance of the model, which indicates that the combination of Attention-Adapter and Res-Adapter can get suitable information on lung sound tasks from large-scale pre-trained models. However, We notice that the performance degradation of methods without Attention-Adapter is even more obvious, which shows the multi-scale cognitive abilities of the adapters.

To verify that our proposed method is also applicable to other transformer-based large-scale pre-trained models, we also experiment with raw SSAST [34] and AST. As shown in Figure 2, we find that our method can achieve competitive or even better scores than full fine-tuning with substantially fewer tunable parameters on the ICBHI dataset. This shows efficacy and robustness when applying our method to a different transformer-based architecture.

4. Conclusion

In this paper, we presented the LungAdapter method, which effectively enhances the efficiency and performance of audio Transformer fine-tuning in lung sound tasks. By adding the serial and parallel adapters to the attention and MLP blocks, the LungAdapter can efficiently capture the distributions of features suitable for lung sounds from pre-trained models and balance the performance and overheads. Experiments demonstrate the compelling performance of our approach with a drastically reduced parameter size. In the field of large models, full fine-tuning is no longer the optimal choice for downstream tasks. We believe that our method can be helpful in different datasets and bring performance breakthroughs on more tasks.

Acknowledge. This work was supported in part by the National Nature Science Foundation of China (No. 62071342, No.62171326) and the Hubei Province Technological Innovation Major Project (No. 2022BCA041).

5. References

[1] A. Gurung, C. G. Scrafford, J. M. Tielsch, O. S. Levine, and W. Checkley, "Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis," *Respiratory medicine*, vol. 105, no. 9, pp. 1396–1403, 2011.

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

[3] Y. Shi, Y. Li, M. Cai, and X. D. Zhang, "A lung sound category recognition method based on wavelet decomposition and bp neural network," *International journal of biological sciences*, vol. 15, no. 1, p. 195, 2019.

[4] J. Heitmann, A. Glangetas, J. Doenz, J. Dervaux, D. M. Shama, D. H. Garcia, M. R. Benissa, A. Cantais, A. Perez, D. Müller *et al.*, "Deepbreath—automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries," *NPJ digital medicine*, vol. 6, no. 1, p. 104, 2023.

[5] B. 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 *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*. Springer, 2018, pp. 33–37.

[6] Y. Ma, X. Xu, and Y. Li, "Lungrn+ nl: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation." in *Interspeech*, 2020, pp. 2902–2906.

[7] Y. Chang, Z. Ren, T. Nguyen, W. Nejdl, and B. Schuller, "Example-based explanations with adversarial attacks for respiratory sound analysis," in *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*, 2022, pp. 4003–4007.

[8] J. Li, J. Yuan, H. Wang, S. Liu, Q. Guo, Y. Ma, Y. Li, L. Zhao, and G. Wang, "Lungattn: advanced lung sound classification using attention mechanism with dual tqwt and triple stft spectrogram," *Physiological Measurement*, vol. 42, no. 10, p. 105006, 2021.

[9] Z. Ren, T. T. Nguyen, and W. Nejdl, "Prototype learning for interpretable respiratory sound analysis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9087–9091.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] S. Gairola, F. Tom, N. Kwatra, and M. Jain, "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[13] S. Bae, J.-W. Kim, W.-Y. Cho, H. Baek, S. Son, B. Lee, C. Ha, K. Tae, S. Kim, and S.-Y. Yun, "Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification," in *Proc. INTERSPEECH 2023*, 2023, pp. 5436–5440.

[14] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[15] J.-W. Kim, S. Bae, W.-Y. Cho, B. Lee, and H.-Y. Jung, "Stethoscope-guided supervised contrastive learning for cross-domain adaptation on respiratory sound classification," *arXiv preprint arXiv:2312.09603*, 2023.

[16] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "Adapterfusion: Non-destructive task composition for transfer learning," *arXiv preprint arXiv:2005.00247*, 2020.

[17] H. Le, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Lightweight adapter tuning for multilingual speech translation," in *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021.

[18] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023.

[19] M. Wang, J. Xing, J. Mei, Y. Liu, and Y. Jiang, "Actionclip: Adapting language-image pretrained models for video action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[20] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 664–16 678, 2022.

[21] Y.-C. Liu, C.-Y. Ma, J. Tian, Z. He, and Z. Kira, "Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 889–36 901, 2022.

[22] D. Yin, L. H. B. Li, and Y. Zhang, "Adapter is all you need for tuning visual tasks," *arXiv preprint arXiv:2311.15010*, 2023.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[25] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[26] L. 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, no. 8, pp. 2938–2947, 2021.

[27] Y. Chang, Z. Ren, T. T. Nguyen, W. Nejdl, and B. W. Schuller, "Example-based Explanations with Adversarial Attacks for Respiratory Sound Analysis," in *Proc. Interspeech 2022*, 2022, pp. 4003–4007.

[28] Z. Wang and Z. Wang, "A domain transfer based data augmentation method for automated respiratory classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9017–9021.

[29] 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.

[30] I. Moummad and N. Farrugia, "Supervised contrastive learning for respiratory sound classification," *arXiv preprint arXiv:2210.16192*, 2022.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[32] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.