ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414385

CONTRASTIVE EMBEDDIND LEARNING METHOD FOR RESPIRATORY SOUND CLASSIFICATION

Wenjie Song, Jiqing Han, Hongwei Song

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, PRC

ABSTRACT

Respiratory sound classification refers to identifying adventitious sounds from given recordings automatically. Due to the difficulty of collection and the expensive manual annotation, there are only limited samples available, which impacts on learning better models. Meanwhile, a majority of these models do not explicitly encourage intra-class compactness and inter-class separability between the learned embeddings, leading to the difficulty of identifying several samples and a reduced generalization performance. To address the problems, we propose a contrastive embedding learning method, where the input is a contrastive tuple. And the composite input strategy provides more possible network inputs. By the comparison among the samples in the tuple, we can learn the slight differences among the similar samples, and the easily-confused samples are more likely to be identified. In the embedding space, we explicitly promote the intra-class compactness and inter-class separability, thereby the generalization performance is improved. Our method is evaluated on ICBHI 2017, and the classification score is increased from 75.61% of a conventional cross-entropy network to 78.18%, outperforming the state-of-the-art methods.

Index Terms— Respiratory sound classification, contrastive learning, CNN

1. INTRODUCTION

Respiratory sound classification means automatically identifying a recording, usually a respiratory cycle, as normal or containing adventitious sounds. It can overcome the subjective problem of conventional auscultation and has the potential for real-time monitoring.

Recently, deep learning has gained a lot of attention due to its unparalleled success in respiratory sound analysis. Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN) and their variants or hybrids are widely used. In [1], a weighted cross-entropy based bi-ResNet with two critical features, i.e. short-time Fourier transform (STFT) and wavelet analysis, was employed to handle the classification task. A CNN model was utilized to compress the spectrograms into deep embeddings [2], which were next regarded as the input of a Linear Discriminant Analysis (LDA) classifier. In [3], a hybrid CNN-RNN model trained with Mel-spectrograms was proposed, and it produced a higher score than VGG16 and MobileNet. A Gated Recurrent Unit (GRU)-based architecture called Noise Masking Recurrent Neural Network (NMRNN) was designed in [4], and it only exploited important respiratory-like Mel-Frequency Cepstral Coefficients (MFCC) frames without redundant noise. MFCC-based feature extraction and advanced Long Short-Term Memory (LSTM) models were integrated and obtained the highest classification score in [5].

Although the performance has been improved by applying deep learning technique, the respiratory sound classification still suffers from limited training samples, since the collection involves in ethics problem and the manual annotation is expensive. In addition, the most commonly used cross-entropy models do not explicitly encourage intra-class compactness and inter-class separability between the learned embeddings [6, 7, 8], resulting in the difficulty of identifying several samples and a reduced generalization performance.

To address the above problems, we propose a contrastive embedding learning method, where the input is a contrastive tuple. Despite the respiratory sounds are still limited, the composite strategy, i.e. the tuple is composed of multiple samples from different categories, provides quite more available network inputs. By the comparison among the samples in the tuple, we can learn the differences among the categories as well as the slight differences among the similar samples. Thus, it is more likely to identify the easily-confused samples. In the embedding space, we explicitly promote the intra-class compactness and inter-class separability, thereby the generalization performance is improved.

We conducted experiments on ICBHI 2017 [9], and verified that our proposed method can learn more compact and well separated embeddings via the limited samples. It is less sensitive to a range of hyperparameters and more stable than other widely used models.

2. METHOD

2.1. The proposed framework

Our framework consists of two networks, namely encoder and classifier, and a contrastive loss, as shown below:

1) The encoder $f(\cdot)$:

Let $\boldsymbol{x} \in \mathcal{X}$ be an acoustic feature, namely anchor, and $y \in \{1, \cdots, L\}$ be its category label. We use $\boldsymbol{x}^+$ and $\{\boldsymbol{x}_i\}_{i=1}^N$ to denote its positive and negatives, i.e. $\boldsymbol{x}$ and $\boldsymbol{x}^+$ are from the same class and $\boldsymbol{x}_i$ is from different classes. The encoder $f(\cdot) : \mathcal{X} \to \mathbb{R}^K$ takes $\boldsymbol{x}$ to generate a K-dimension embedding $h = f(\boldsymbol{x})$, which inherits all superscripts and subscripts from $\boldsymbol{x}$. Our framework allows various choices of the network architecture without any constraints. We choose simplicity and adopt a CNN as the encoder where the output after the global average pooling layer is used as the embedding.

2) Contrastive loss:

Consider a tuple of training samples $\{\boldsymbol{x}, \boldsymbol{x}^+, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, the corresponding embeddings generated by the encoder are $\{\boldsymbol{h}, \boldsymbol{h}^+, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_N\}$. Previous works have indicated that normalizing the embeddings to the unit hypersphere in $\mathbb{R}^K$ always improves performance [10]. Thus, we instead use the normalized embeddings, marked as $\{\boldsymbol{z}, \boldsymbol{z}^+, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_N\}$, to define our contrastive loss as follows:

$$\mathcal{L} = -\log \frac{\exp\left(\boldsymbol{z}^\top \boldsymbol{z}^+\right)}{\exp\left(\boldsymbol{z}^\top \boldsymbol{z}^+\right) + \sum_{k=1}^N \exp\left(\boldsymbol{z}^\top \boldsymbol{z}_k\right)} \tag{1}$$

where $\boldsymbol{z}^\top \boldsymbol{z}^+$ computes the inner product between the normalized vectors $\boldsymbol{z}$ and $\boldsymbol{z}^+$ (i.e. cosine similarity) in the unit hypersphere, so does $\boldsymbol{z}^\top \boldsymbol{z}_k$ $(k = 1 \cdots N)$.

For any anchor $\boldsymbol{z}$, it forces the encoder to maximize the cosine similarity between $\boldsymbol{z}$ and its positive $\boldsymbol{z}^+$, while simultaneously minimize the ones between $\boldsymbol{z}$ and all its negatives $\boldsymbol{z}_k$ $(k = 1 \cdots N)$. Thus, the embeddings which are compact in intra-class and well separated in inter-class are obtained. On the other hand, it can be interpreted as a binary classification loss, that is identifying the positive from the N negatives. To minimize it, we need to learn the differences among the categories as well as the slight differences among the samples simultaneously. Thus, the easily-confused samples can be identified more easily.

3) The classifier $g(\cdot)$:

The classifier $g(\cdot) : \mathbb{R}^K \to \mathbb{R}^L$ maps the learned embeddings into specific categories. Since the embeddings are compact in intra-class and well separated in inter-class, we can use a simple linear classifier $\boldsymbol{W}$ to recognize them, namely $\hat{y} = g(\boldsymbol{h}) = \sigma\left(\boldsymbol{W}^\top \boldsymbol{h}\right)$, where σ is a softmax function.

To accelerate convergence, we train the encoder from a pre-trained network instead of randomly initialized parameters. When training the model, we first choose the positive and negatives, according to their category labels, to update the encoder with Eq.(1). Then we use the learned embeddings to train the classifier. It should be highlighted that the encoder is frozen when training the classifier. While obviously, the gradient cannot be propagated to the classifier when updating the encoder. It means that the two stages of our framework are independent and alternatively optimized until convergence.

2.2. Superiority of our contrastive loss

The concept of contrast has been proposed for years and there are already several other contrastive objectives in literatures. Although yielding promising progress, these works, such as triplet loss [10], logistic loss [11], often suffer from slow convergence and poor local optima [12], since they do not weigh the negatives by their relative hardness [13].

In our method, the gradient of contrastive loss $\mathcal{L}$ to the learned embedding $\boldsymbol{h}$ can be expressed as:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}} = \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}}\right|_{\text{pos}} + \sum\nolimits_{k=1}^N \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}}\right|_k \tag{2}$$

It consists of two components produced by the positive and the N negatives respectively, and the specific formulas are:

$$\left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}}\right|_{\text{pos}} \propto \left(\left(\boldsymbol{z}^\top \boldsymbol{z}^+\right) \cdot \boldsymbol{z} - \boldsymbol{z}^+\right)\left(1 - P^+\right) \tag{3}$$

$$\left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}}\right|_k \propto \left(\boldsymbol{z}_k - \left(\boldsymbol{z}^\top \boldsymbol{z}_k\right) \cdot \boldsymbol{z}\right) \cdot P_k \tag{4}$$

where $\boldsymbol{z}$ is the normalized vector of $\boldsymbol{h}$, which inherits all superscripts and subscripts from $\boldsymbol{h}$, and we define $P^+ = \frac{\exp\left(\boldsymbol{z}^\top \boldsymbol{z}^+\right)}{\exp\left(\boldsymbol{z}^\top \boldsymbol{z}^+\right) + \sum_{k=1}^N \exp\left(\boldsymbol{z}^\top \boldsymbol{z}_k\right)}$, $P_k = \frac{\exp\left(\boldsymbol{z}^\top \boldsymbol{z}_k\right)}{\exp\left(\boldsymbol{z}^\top \boldsymbol{z}^+\right) + \sum_{k=1}^N \exp\left(\boldsymbol{z}^\top \boldsymbol{z}_k\right)}$ $(k = 1 \cdots N)$ for simplicity.

For an easy negative, which is already far away from the anchor, $\boldsymbol{z}^\top \boldsymbol{z}_k \approx -1$ and P_k is small. Thus:

$$\left(\boldsymbol{z}_k - \left(\boldsymbol{z}^\top \boldsymbol{z}_k\right) \cdot \boldsymbol{z}\right) \cdot P_k \approx \left(\boldsymbol{z}_k + \boldsymbol{z}\right) \cdot P_k \approx 0 \tag{5}$$

While for a hard one, i.e. a negative that is still close to the anchor, $\boldsymbol{z}^\top \boldsymbol{z}_k \approx 0$ and P_k is moderate, then

$$\left(\boldsymbol{z}_k - \left(\boldsymbol{z}^\top \boldsymbol{z}_k\right) \cdot \boldsymbol{z}\right) \cdot P_k \approx \boldsymbol{z}_k \cdot P_k > 0 \tag{6}$$

Thus, for an easy negative, its contribution to the gradient is small, while for a hard one, it is large. Our contrastive loss effectively weighs different negatives and helps the model focus more on hard ones. Thereby the slow convergence and other problems can be alleviated.

2.3. An effective batch construction

Suppose we directly input a $(N+2)$-tuple to the encoder. When the batch size of Stochastic Gradient Descent (SGD) [14] or other optimizer is M, there are $M \times (N+2)$ samples to be passed through f at one update. Limited by the scale of training samples, one may be evaluated repeatedly as anchor, positive or negative in a same batch, which means redundant and impractical for a deep neural network. More importantly, it is difficult to collect a relatively balanced dataset since the incidences of different adventitious sounds vary a lot. The category distribution of negatives is inconsistent with the anchor and the positive, resulting in batch normalization

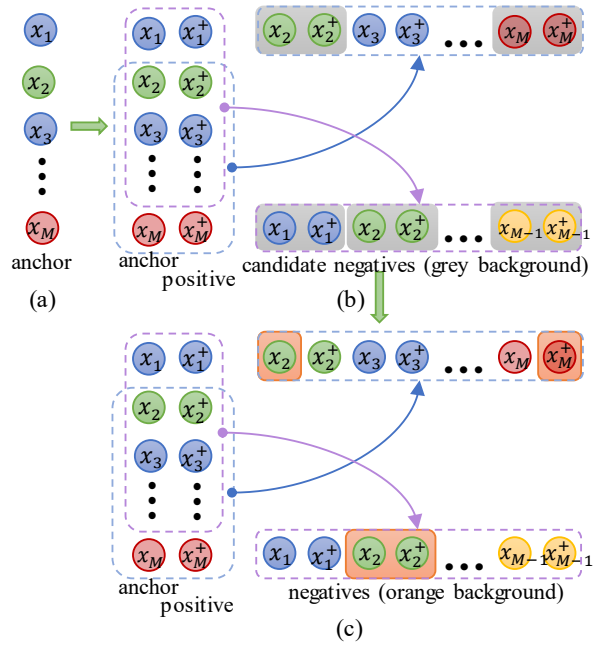

Fig. 1. An illustration of our batch construction.

unavailable. Thus, we introduce an effective batch construction approach as shown in Fig. 1

We first randomly select a batch of samples $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M\}$ as the queue of anchor without any constrains. According to their category labels, we contruct a corresponding positive queue $\{\boldsymbol{x}_1{}^+, \cdots, \boldsymbol{x}_M{}^+\}$, where $\boldsymbol{x}_i{}^+ \ (i = 1 \cdots M)$ is a random choice of the samples with same label as $\boldsymbol{x}_i$. When considering the negatives, all samples belonging to other classes in the two queues can be candidates. And we reserve N of them to ensure the number of negatives is consistent.

Despite our contrastive loss shares similar idea with N-pair loss [12] and supervised contrastive loss [15], we would like to point out that we have differences in many aspects. Firstly, we use a pre-trained network, which requires significantly fewer epochs and achieves a better performance, instead of training from randomly initialized parameters. Secondly, our batch construction approach is specially designed for respiratory sound classification. N-pair loss demands N pairs of samples from N different categories, and the N is usually much greater than the category number defined in medicine. Supervised contrastive loss can be reasonable only when the data distribution is balanced like ImageNet. And worse, for the categories with very few samples, it cannot guarantee the existence of legal positives, resulting in training errors. Thus, the two methods cannot be directly applied to our task. Most importantly, the problems we want to address are different: N-pair loss was proposed for slow convergence caused by triplet loss, supervised contrastive loss was directly inspired by self-supervised learning, aiming at the shortcomings of cross-entropy, such as poor margin and lack of robustness to noisy labels. While we are more focused on the composite input strategy, which can alleviate the data sparsity, and learning more discriminative embeddings by comparison among the samples simultaneously.

Table 1. The network architecture of the encoder

Name	Parameters	Output size
Input	–	64×320
Conv1	$conv\ [3 \times 3, 32]$ $conv\ [3 \times 3, 32]$	$64 \times 320 \times 32$
Pool1	$maxpool\ [2 \times 2]$, $stride\ 2$	$32 \times 160 \times 32$
Conv2	$conv\ [3 \times 3, 64]$ $conv\ [3 \times 3, 64]$	$32 \times 160 \times 64$
Pool2	$maxpool\ [2 \times 2]$, $stride\ 2$	$16 \times 80 \times 64$
Conv3	$conv\ [3 \times 3, 128]$ $conv\ [3 \times 3, 128]$	$16 \times 80 \times 128$
Pool3	$maxpool\ [2 \times 2]$, $stride\ 2$	$8 \times 40 \times 128$
Conv4	$conv\ [3 \times 3, 256]$ $conv\ [3 \times 3, 256]$	$8 \times 40 \times 256$
Pool4	global average pool	256

3. EXPERIMENTS

3.1. Dataset and evaluation metrics

The ICBHI 2017 used in this paper is the largest publicly available respiratory sound dataset. It consists of a total of 5.5 hours of recordings containing 6898 respiratory cycles, of which 1864 contain crackles, 886 contain wheezes, 506 contain both crackles and wheezes and the rest are normal, in 920 annotated audio samples from 126 subjects [9]. To evaluate the effectiveness of our method, we divide the dataset into training (80%) and test (20%) set, and 20% of the training samples further constitute the validation set. The standard benchmarks, i.e. Accuracy (Acc.), Sensitivity (Sen.), Specificity (Spe.) and ICBHI score are used [16].

3.2. Experimental setup

1) Feature extraction and data augmentation

We first resample all audio recordings, which were stored at a variety of sample rates, to 16kHz and segment them to respiratory circles according to the provided onsets and offsets. Then we convert the circles to 64-dimension log Mel-spectrograms with a window size of 1024 over a 256-sample hop. Since their durations vary a lot, we repeat the short ones and trim off partially the long ones to ensure the shapes of the features are unified (64x320).

The effectiveness of data augmentation in respiratory sound classification has been proved in [3, 17]. Hence we

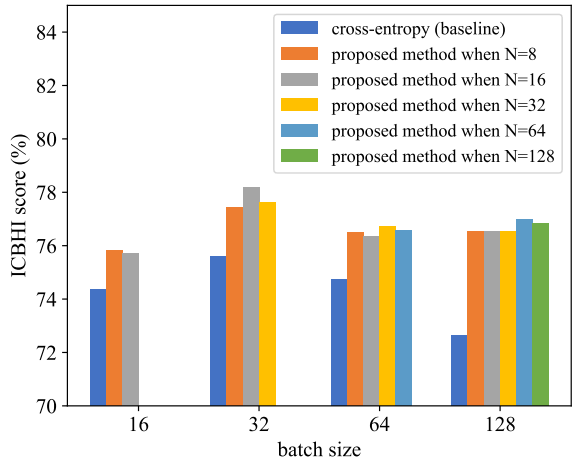

Fig. 2. The impact of batch size and number of negatives.

adopt 4 common strategies, namely white noise adding, time shifting, time stretching and pitch shifting.

2) Networks architecture

A simple CNN, where batch normalization and Rectified Linear Unit (ReLU) are applied to the output of every concolutional layer, is used as the encoder and its architecture is represented in Table 1. As we mentioned earlier, the classifier $\boldsymbol{W} \in \mathbb{R}^{256\times4}$ is linear, softmax and cross-entropy are used respectively as its activation function and objective function. The baseline is also a CNN whose architecture is a concatenation of the encoder and classifier. All the models are trained using an Adam optimizer. An initial learning rate of 1e-3 and maximum iterations of 80 are set for the baseline and the pretraining of encoder, 1e-5 and 30 for our contrastive method.

3.3. Experimental results

1) Influence of batch size and number of negatives

Our contrastive loss demands the anchor to be far away from the N negatives simultaneously, and the learning difficulty will gradually increase as N becomes bigger, thereby affecting the quality of learned embeddings. On the other hand, batch size plays an important role in our method. It not only directly influences the initial weights of the encoder, but also determines the upper bound of N in our batch construction. Thus, we evaluate our method with different batch sizes and quantities of negatives, and compare it with conventional cross-entropy method (i.e. the baseline), as shown Fig. 2.

It can be observed that our method can improve the performance under a variety of batch sizes and quantities of negatives. The best ICBHI score (78.18%) is achieved when batch size is 32 and N is 16. Not surprisingly, the scores with a batch size of 32 wholly outperform the others, since a proper batch size provides better initial weights of the encoder, which usually means a greater possibility to converge to a better local minimum. In addition, compared with the baseline, whose performance drops when the batch size continues increasing inappropriately (greater than 32), our method is less sensitive to a range of batch size. Even when the batch size is 128, the baseline decreases below 73%, it still guarantees stable scores over 76.5%. When fixing the batch size, N does not seem to be very influential. We attribute it to the small-scale of our dataset, and the diversity of sampled negatives is insufficient even when N increases.

Table 2. Comparison between the proposed method and state-of-the-art methods.

Method	Acc. (%)	Sen. (%)	Spe. (%)	score (%)
STFT + wavelet Bi-ResNet [1]	67.44	58.54	80.06	69.30
STFT parallel-pooling CNN [2]	71.15	57.67	83.24	70.45
MFCC LSTM [5]	-	62.00	85.00	74.00
Mel-spectrogram CNN (baseline)	76.44	64.69	**86.53**	75.61
Ours	**78.73**	**70.93**	85.44	**78.18**

2) Comparison with state-of-the-art systems

When comparing our proposed method with some existing works, for fairness, we ensure the same training/test splitting is used and obtain the results, as shown in Table 2. Since the baseline tends to classify a sample as the most common normal sound, its specificity is the highest. While on the whole, our method significantly outperforms the others.

4. CONCLUSION

In this paper, we propose a contrastive embedding learning method for respiratory sound classification. Aiming at data sparsity, we are the first to use the composite input strategy to obtain more available network inputs. And it indicates that data augmentation is not the only way to solve the problem. In addition, we introduce the contrastive learning into the task. By the comparison among the samples, we can obtain more discriminative embeddings and improve the generalization performance. Our successful attempt confirms the possibility of using advanced learning by comparison theory and composite input strategy to learn better embeddings from the same limited samples.

5. ACKNOWLEDGEMENTS

This research is supported by the National Natural Science Foundation of China under grant No. U1736210.

6. REFERENCES

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

[2] Fatih 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, 2020.

[3] Jyotibdha 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, pp. 535–544, 2020.

[4] Kirill Kochetov, Evgeny Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *ICANN*, 2018.

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

[6] Weiyang Liu, Y. Wen, Zhiding Yu, and Meng Yang, "Large-margin softmax loss for convolutional neural networks," in *ICML*, 2016.

[7] Bo Dai and D. Lin, "Contrastive learning for image captioning," in *NIPS*, 2017.

[8] Gamaleldin F. Elsayed, Dilip Krishnan, H. Mobahi, Kevin Regan, and S. Bengio, "Large margin deep networks for classification," *ArXiv*, vol. abs/1803.05598, 2018.

[9] Bruno Rocha, Dimitris Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, Ana Oliveira, C. Jácome, A. Marques, R. Paiva, Ioanna Chouvarda, P. Carvalho, and N. Maglaveras, "A respiratory sound database for the development of automated classification," in *BHI 2017*, 2017.

[10] Florian Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015.

[11] Tomas Mikolov, Kai Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[12] Kihyuk Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NIPS*, 2016.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," *ArXiv*, vol. abs/2002.05709, 2020.

[14] D. Rumelhart, Geoffrey E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," 1986.

[15] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, A. Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *ArXiv*, vol. abs/2004.11362, 2020.

[16] B. M. Rocha, D. Filos, L. Mendes, Gorkem Serbes, Sezer Ulukaya, Y. Kahya, N. Jakovljevic, T. L. Turukalo, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, N. Maglaveras, Rui Pedro Paiva, Ioanna Chouvarda, and P. de Carvalho, "An open access database for the evaluation of respiratory sound classification algorithms.," *Physiological measurement*, vol. 40 3, pp. 035001, 2019.

[17] J. Than, N. M. Noor, O. M. Rijal, R. M. Kassim, and A. Yunus, "Advanced and minor lung disease severity classification using deep features," *2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 122–127, 2019.