LungAttn: Advanced Lung Sound Classification Using Attention Mechanism with Dual TQWT and Triple STFT Spectrogram

Jizuo Li[1§], Jiajun Yuan[2,3,5,6§], Hansong Wang[2,5], Shijian Liu[2,5], Qianyu Guo[1], Yi Ma[1], Yongfu Li[1]*, Liebin Zhao[4,5]*, Guoxing Wang[1]

[§]Jizuo Li and Jiajun Yuan contributed equally to this work.

*Yongfu Li and Liebin Zhao are co-corresponding authors for this work.

[1]Department of Micro-Nano Electronics and MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, 200240, China

[2]Pediatric AI Clinical Application and Research Center, Shanghai Children's Medical Center, and Child Health Advocacy Institute, China Hospital Development Institute of Shanghai Jiao Tong University

[3]School of Computer Engineering and Science, Shanghai University, Shanghai, China

[4]Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine

[5]Shanghai Engineering Research Center of Intelligence Pediatrics (SERCIP)

[6]Sanya Maternity and Child Care Hospital

E-mail: `yongfu.li@sjtu.edu.cn, zhaoliebin@126.com`

Aug 2021

Abstract. *Objective.* Auscultation of lung sound plays an important role in the early diagnosis of lung diseases. This work aims to develop an automated adventitious lung sound detection method to reduce the workload of physicians. *Approach.* We propose a deep learning architecture, LungAttn, which incorporates augmented attention convolution into ResNet block to improve the classification accuracy of lung sound. We adopt a feature extraction method based on dual tunable Q-factor wavelet transform (TQWT) and triple short-time Fourier transform (STFT) to obtain a multi-channel spectrogram. Mixup method is introduced to augment adventitious lung sound recordings to address the imbalance dataset problem. *Main results.* Based on the ICBHI 2017 challenge dataset, we implement our framework and compare with the state-of-the-art works. Experimental results show that LungAttn has achieved the *Sensitivity,* S_e, *Specificity,* S_p and *Score* of 36.36%, 71.44% and 53.90%, respectively. Of which, our work has improved the *Score* by 1.69% compared to the state-of-the-art models based on the official ICBHI 2017 dataset splitting method. *Significance.* Multi-channel spectrogram based on different oscillatory behavior of adventitious lung sound provides necessary information of lung sound recordings. Attention mechanism is introduced to lung sound classification methods and has proved to be effective. The proposed LungAttn model can potentially improve the speed and accuracy of lung sound classification in clinical practice.

Keywords: Lung Sound Classification, Crackle and Wheeze, Feature Extraction, Augmented Attention Convolution

1. Introduction

According to the World Health Organization, respiratory diseases, such as pneumonia, asthma, bronchitis, lung cancer, and chronic obstructive pulmonary disease (COPD), are one of the most common mortality factors in the world, causing the death of more than 3 million people each year worldwide (Cukic et al. 2012). These respiratory diseases have a direct impact on people's social, economic, and health life. Early diagnosis is the key factor for preventing the spread of respiratory diseases and limiting the adverse effects on people's life.

Lung sound characteristics provide important information in the diagnosis of respiratory diseases (Marques and Anne 2009; Aziz et al. 2019). The two most common adventitious lung sounds are crackle and wheeze (Rocha, Cristina, and Alda 2016). Wheeze sound is similar to a musical high-pitched continuous sound, and it is caused by airway narrowing or airflow limitation, which is an important symptom of asthma and COPD (Pramono, Imtiaz, and Rodriguez-Villegas 2019). Crackle is an explosive and discontinuous sound during the inspiratory and expiratory process of the breathing cycle, which is associated with obstructive airway diseases and interstitial lung diseases, such as chronic bronchitis, pneumonia, and lung fibrosis (Piirila and Sovijarvi 1995). The presence of crackle and wheeze adventitious sound indicates the presence of corresponding respiratory diseases.

Auscultation is a technique where specialists use a stethoscope to detect adventitious lung sounds and identify the possible lung diseases (Sengupta, Sahidullah, and Saha 2016). It provides a simple, low-cost, and non-invasive method for respiratory disease diagnosis. However, the auscultation technique based on stethoscope suffers from two disadvantages. Firstly, auscultation requires expert physicians to analyze the lung sounds. In impoverished environments with a shortage of expert physicians, any sudden and massive infectious respiratory diseases outbreak (such as the pneumonia complication from the coronavirus (Xie et al. 2020)) can further exacerbate the spread of the diseases and the increase of death rate. Secondly, even if the patients are diagnosed by experienced physicians, there might be subjectivity in the interpretations of lung sounds, causing inter-listener variability (Lia et al. 2016). Therefore, it is necessary to develop an automatic respiratory detection method to reduce the workload for physicians and eliminate subjectivity during diagnosis.

To accelerate the development of automatic respiratory detection methods, in 2017, the International Conference on Biomedical and Health Informatics (ICBHI) released an open source dataset and organized a scientific challenge to classify adventitious lung sounds(Rocha et al. 2019). The publicly available ICBHI 2017 dataset has inspired many research groups to explore the problem using various classification methods. Compared to the earlier respiratory lung sound classification algorithms (Zhang et al.

2015; Liu et al. 2015; Wiśniewski and Zieliński 2015; Lozano, Fiz, and Jané 2016), there are significantly newer developments that have been proposed within these few years (Jakovljević and Lončar-Turukalo 2018; Chambres, Hanna, and Desainte 2018; Serbes, Ulukaya, and Kahya 2018; Ma et al. 2019; Ma, Xu, and Li 2020; Pham et al. 2021). There are several improvements made in (1) feature extractor, and (2) classifier within the automatic respiratory detection systems.

Feature extractor extracts various features of respiratory signals in the time domain and frequency domain, while a classifier is used to analyze these features and classify different types of lung sound. Feature extraction methods can be categorized into statistical feature extraction and spectrogram extraction. Jakovljević and Lončar-Turukalo (2018) extracted statistical features based on Mel-Frequency Cepstral Coefficient (MFCC). Chambres, Hanna, and Desainte (2018) computed statistical spectral information like barkbands, energybands, melbands, etc. In (Serbes, Ulukaya, and Kahya 2018), mathematical features including mean, skewness, kurtosis, standard deviation, and energy are calculated after wavelet transform. Chen et al. (2019) proposed optimized S-transform to obtain spectrograms. Acharya and Basu (2020), Ma et al. (2019), Ma et al. (2019), and Pham et al. (2021) have extracted spectrograms such as Mel-spectrogram, short-time Fourier transform (STFT) spectrogram, stacked wavelet coefficient spectrogram, and Gammatone spectrogram. However, less attention has been paid to extract separate features according to the nature of these adventitious lung sounds.

Classifiers are grouped into traditional machine learning classifiers and deep learning classifiers. Jakovljević and Lončar-Turukalo (2018) adopted hidden Markov model with Gaussian mixture model (HMM-GMM) to analyze the statistical MFCC of lung sounds. Chambres, Hanna, and Desainte (2018) and Serbes, Ulukaya, and Kahya (2018) adopted traditional machine learning methods of boosted decision tree and SVM to classify lung sounds. Kochetov et al. (2018) proposed to use a deep recurrent neural network with a noise-masking model to process respiratory signals. Acharya and Basu (2020) and Messner et al. (2020) proposed to use hybrid CNN-RNN models to perform classification. In (Ma et al. 2019; Ma et al. 2019), a LungRBN model based on ResNet has been developed. A CNN model is designed in (Shuvo et al. 2020) to detect respiratory diseases using hybrid scalogram features. Demir, Ismael, and Sengur (2020) developed a CNN model using parallel pooling structure. However, most of the existing works did not address the severely imbalance problem, which limits the performance of existing classifiers.

To overcome these challenges, we propose an advanced adventitious lung sound classification method, LungAttn, and the key contributions of our works are summarized as follows:

(i) **Feature Extraction**: According to different oscillatory and duration characteristics of crackle and wheeze, we propose dual tunable Q-factor wavelet transform (TQWT) (Selesnick 2011a) to perform resonance-based decomposition followed by triple STFT to obtain separate spectrograms including the spectrogram of tran-

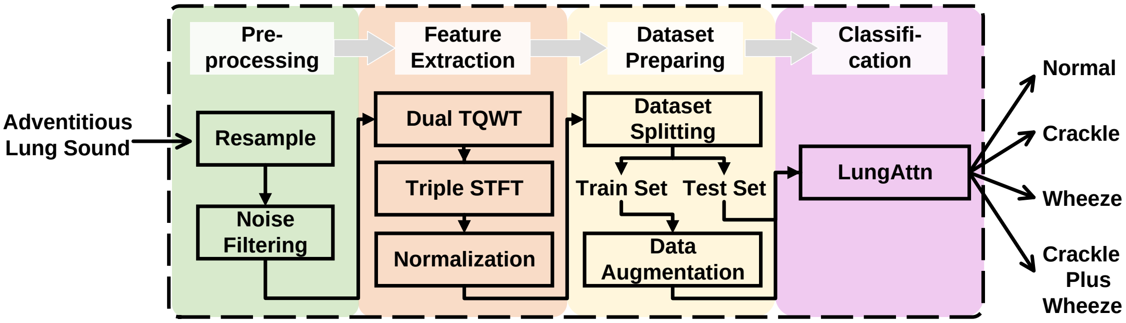


Figure 1. An conceptual idea of proposed framework.

sient, oscillatory and residual components, corresponding to crackle, wheeze, and noise characteristics, respectively.

(ii) **Data Augmentation**: Based on the classification requirement in the ICBHI 2017 dataset, we propose to use mixup (Zhang and Cisse 2018) to generate new samples of crackles and wheezes by combining two original samples, increasing the percentage of adventitious lung sound samples from 47.2% to 71.3%.

(iii) **Classification Model**: We introduce attention augmented convolution (Bello et al. 2019) in the ResNet architecture to capture long spatial and temporal intervals relationship in the spectrogram.

(iv) **Experimental Results**: We conduct experiments to demonstrate the effectiveness of our proposed feature extraction method, data augmentation method, and classification model. Based on the ICBHI 2017 dataset, our proposed LungAttn model outperforms state-of-the-art works (Jakovljević and Lončar-Turukalo 2018; Chambres, Hanna, and Desainte 2018; Serbes, Ulukaya, and Kahya 2018; Ma et al. 2019; Ma, Xu, and Li 2020; Pham et al. 2021).

The rest of this paper is organized as follows. Section 2 provides details of our proposed pre-processing method, feature extraction techniques, and mixup data augmentation. Section 3 presents our proposed LungAttn framework. Section 4 states the experimental setup and evaluation method. Section 5 discusses the effectiveness of our proposed techniques followed by the comparison with state-of-the-art works. We conclude our work in Section 6.

2. Dataset Preprocessing and Feature extraction

2.1. ICBHI 2017 Dataset

ICBHI 2017 dataset is the largest annotated publicly available benchmark for respiratory sound classification, collected from 126 subjects with a total duration of 5.5 hours (Rocha et al. 2019). In this dataset, a record is defined as the lung sound collected from one patient and a cycle is defined as a respiratory cycle from a patient. Each breathing

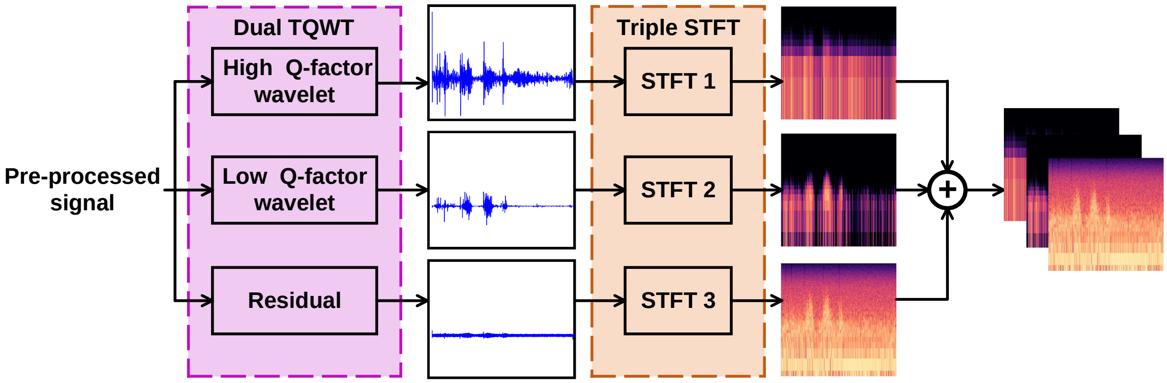


Figure 2. The proposed feature extraction method using dual TQWT and triple STFT.

cycle in a recording is annotated by an expert as one of the four classes: "Normal", "Crackle", "Wheeze", or "Crackle plus Wheeze". The dataset includes a total of 6,898 different respiratory cycles with 3,642 "Normal" cycles, 1864 "Crackle", 886 "Wheeze", and 506 "Crackle plus Wheeze" cycles. Since these recordings were collected from various hospitals using different recording devices, the sampling frequency, background noise level, and the number of respiratory cycles vary among patients.

2.2. Pre-processing Techniques

For healthy vesicular breathing lung sounds, the vast majority of the power spectra falls in the range of 50-to-1,000 Hz, while the frequency range of adventitious respiratory sounds, such as wheeze and crackle extends beyond 50-to-2,000 Hz (Perna and Tagarelli 2019). Since the sampling frequency of audio recordings ranges from 4,000-to-44,100 Hz, the signals are resampled to 4,000 Hz to allow a simple and consistent processing of these signals. After resampling to 4kHz, the maximum frequency in spectrum is limited to 2kHz. To suppress the environmental noises, such as heartbeat, motion artifacts and audio sound (Bohadana, Izbicki, and Kraman 2014), a 5-th order Butterworth high-pass filter with 2dB maximum attenuation in passband is used to retain the frequency band from 50-to-2,000 Hz. Since the frequency ranges for each type of adventitious sounds are highly overlapping and its exact location is unknown, frequency based decomposition using linear time invariant filters is unable to separate each type. Hence, there is a need to decompose these signals with non-linear filters based on the individual characteristics of these signals.

2.3. Feature Extraction

As illustrated in Fig. 2, our proposed feature extraction method includes (1) dual tunable Q-factor wavelet transform to perform signal decomposition, and (2) triple short-time Fourier transform to convert signals into spetrograms.

Tunable Q-factor Wavelet Transform (TQWT): The TQWT is basically a fully-discrete wavelet transform, which facilitates the analysis of oscillatory signals with easily adjustable parameters. The input of TQWT depends on the Q-factor (Q) and the over-sampling rate (r) to control the oscillatory behavior of the signal (Selesnick 2011b). The TQWT's parameters can be adjusted to match the oscillatory behavior of biomedical signal analysis, allowing a clearer distinction between different classes of signals (Shivnarayan and Ram 2013; Shivnarayan and Trilochan 2017). Crackles are transient waveforms with sudden bursts, it can be represented with the low Q-factor wavelet. Similarly, wheezes tend to have oscillatory waveforms, it can be represented with the high Q-factor wavelet. Thus, based on Morphological Component Analysis (MCA) (Bobin et al. 2007), we propose a dual Q-factor wavelet transform (TQWT), which performs resonance-based decomposition on respiratory sound into three components, namely, transient component (crackle), oscillatory component (wheeze), and residual component (noise).

Short-time Fourier Transform (STFT): After obtaining transient, oscillatory, and residual components of the signal, STFT is applied to each component to obtain time-frequency representation. The frequency of crackle signals locates around 350-650 Hz and its duration is less than 20 ms, while the frequency of wheeze signals ranges from 100-2000 Hz and it lasts for over 80 ms (Serbes, Ulukaya, and Kahya 2018). Due to the non-linear and non-stationary characteristics of crackle and wheeze sounds, a Hanning window function is used to capture frequency domain information within a short time interval. Since the duration of each adventitious sound is different, we apply specific window length for each component. A 20 ms window length with 50% overlap between two adjacent windows is applied on the transient component. For the oscillatory and residual components, we apply STFT with 80 ms and 200 ms window length, respectively with 50% overlapping of the adjacent windows. After applying different STFT conditions on these components, three time-frequency images are obtained, which are resized to 224×224 and concatenated together to form a three-channel spectrogram for the neural network.

For our spectrogram conversion process, recordings with different time durations are directly converted to spectrograms first, and then the spectrograms are resized to the same size. The spectrogram resizing process is detailed in Algorithm 1. The algorithm traverses all elements for the resized spectrogram D and find the corresponding location in the original spectrogram S (Line 3-6). The value of an element in the resized spectrogram D is calculated by four corresponding elements in the original spectrogram S (Line 7). The process does not require cropping or padding operation on original recordings to avoid information loss or redundancy. Fig. 3 visualizes the TQWT waveforms and STFT spectrograms of different lung sound types. Notice that the spectrograms of transient, oscillatory, and residual components share the same time and frequency units for the same recording. However, due to the fixed size of spectrograms, recordings with longer time duration have spectrograms with smaller time resolution. The concatenated three-channel spectrogram is processed with Min-Max normalization

Algorithm 1 Spectrogram Resizing

```
Input: Original spectrogram $S$ with size of $M \times N$; Desired size of spectrogram $W \times H$;
Output: Resized spectrogram $D$;
for $i = 0$ to $W - 1$ do
    for $j = 0$ to $H - 1$ do
        $src_x = (i + 0.5) \times \frac{M}{W} - 0.5$;
        $src_y = (j + 0.5) \times \frac{N}{H} - 0.5$;
        $x = \lfloor src_x \rfloor$, $u = src_x - x$;
        $y = \lfloor src_y \rfloor$, $v = src_y - y$;
        $D(i, j) = (1 - u)(1 - v)S(x, y) + (1 - u)vS(x, y + 1) + u(1 - v)S(x + 1, y) + uvS(x + 1, y + 1)$;
    end for
end for
return $D$;
```

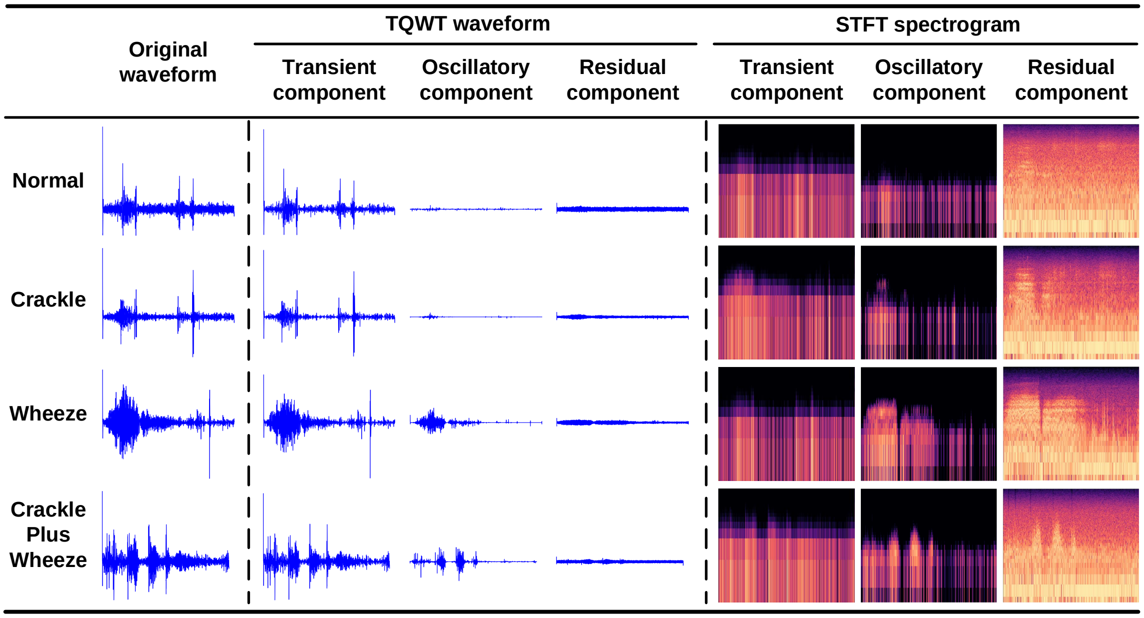


Figure 3. Visualization for original lung sound waveform, TQWT waveform, and STFT spectrogram of "Normal", "Crackle", "Wheeze", and "Crackle plus Wheeze".

to fix all the values in the spectrogram to the range of [0,1].

2.4. Mixup Data Augmentation

As shown in Table 1, the majority class "Normal" and the minority class "Crackle plus Wheeze" have 3,642 samples and 506 samples, respectively, which is about 7.20× different. This is considered an imbalanced dataset, which is practically a common problem in all the medical classification tasks due to the low probability of obtaining

Table 1. Distribution of Classes in ICBHI 2017 and after augmentation method.

	Before augmentation		**After augmentation**	
	Amount	Ratio	Amount[1]	Ratio[2]
Normal	3642	0.528	2060	0.287
Crackle [C]	1864	0.270	2034	0.284
Wheeze [W]	886	0.128	1648	0.230
C plus W	506	0.073	1421	0.199
Total	6898	1	7163	1

1: The amount is number of samples used for training.

2: Average categories ratio after augmentation. It was computed in an epoch.

abnormal samples (Rollins et al. 2015). This problem easily causes the classification model to ignore the minority classes or create over-fitting. Thus, a data augmentation method is needed to address the imbalanced problem (Luis and Jason 2017).

Since adventitious lung sound is feeble and susceptible to different environment noises (Bohadana, Izbicki, and Kraman 2014), traditional data augmentation methods (Ko et al. 2015) generate new samples by changing the characteristic of the signals in the time or frequency domain. However, these methods are not effective in improving the detection accuracy, which has been proven by (Ma, Xu, and Li 2020). Some researches has adopted variational autoencoder to augment a new lung sound dataset (García-Ordás et al. 2020). However, deep learning based data augmentation is computationally costly.

To address the aforementioned problem, we have proposed a mixup method (Zhang and Cisse 2018), which is a data-agnostic data augmentation method that makes decision boundaries transit linearly from class to class and provides a smoother estimate of uncertainty. Fig. 4 illustrates the procedure of using a mixup method, which is based on the understanding of the characteristics of each adventitious lung sound type in ICBHI 2017 dataset. For example, crackle tends to have a short and explosive sound, which can be viewed as a specific sound event with normal lung sound as background. Hence, it is reasonable to combine "Normal" cycles with "Crackle" cycles to increase the number of "Crackle" cycles in the dataset. Similarly, the number of "Wheeze" cycles in the dataset can be increased by combining the "Normal" cycles and "Wheeze" cycles. The new "Crackle plus Wheeze" cycles are obtained by mixing the "Crackle" and "Wheeze" cycles.

To combine two randomly selected samples with their labels in the training dataset and generate a new set of samples and labels, the equations are performed as follows:

$$\begin{aligned} \widetilde{x} &= \lambda x_i + (1-\lambda)x_j, \\ \widetilde{y} &= \lambda y_i + (1-\lambda)y_j, \end{aligned} \tag{1}$$

where (x_i, x_j) are two feature vectors from sample i and j, and (y_i, y_j) are one-hot encoded class labels of the two features. $\lambda \in [0, 1]$ is a random number generated

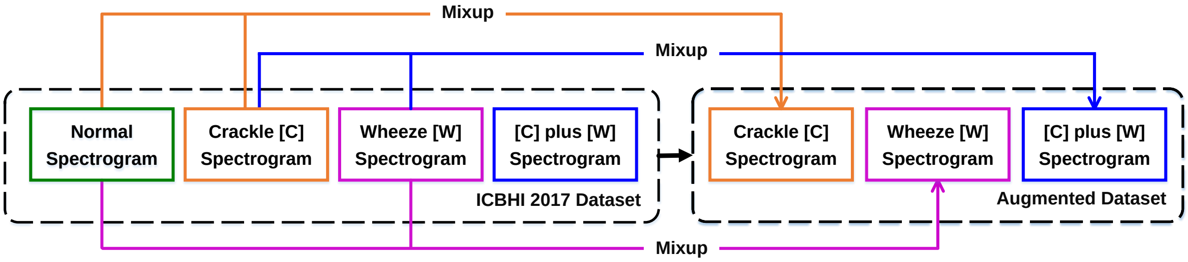


Figure 4. The proposed data augmentation result using the mixup method with ICBHI 2017 dataset.

according to Beta distribution (Goodfellow, Bengio, and Courville 2016).

Although the label is kept in one-hot encoded format for the classification task, it might not be appropriate for the adventitious lung sound classification task. In reality, there is no quantitative standard to diagnose adventitious lung sounds, and medical staff analyzes each sample according to their experience. Furthermore, the severity of the illness varies among patients and it is hard to obtain the identical results for the same type of illness. Therefore, to address the aforementioned problem, we use the probability function to represent the generated label vector with the value varying from 0 to 1, which is determined by λ.

3. LungAttn Architecture

3.1. Attention Augmented Convolution

Convolutional neural networks have shown great success in image classification tasks (Krizhevsky, Sutskever, and Hinton 2012). However, the local nature of the convolutional kernel prevents it from capturing global contexts in an image, often necessary for better classification accuracy (Rabinovich et al. 2007). Thus, attention mechanism is introduced into the convolution neural network, which is a way to consider the nonuniform weights of input feature vectors and optimize the process of learning some target classes (Vaswani et al. 2017). Since the attention mechanism is able to capture long-distance interactions, we have adopted augmented attention convolution in our ResNet-based architecture to measure the relationship of STFT spectrograms across time and frequency domains (Bello et al. 2019).

An image with shape (H, W, F_{in}) is flattened to a matrix $X \in \mathbb{R}^{HW\times F_{in}}$. We define a *query*: $Q = XW_q$, a *key*: $K = XW_k$, and a *value*: $V = XW_v$, where W_q, $W_k \in \mathbb{R}^{F_{in}\times d_k^h}$, and $W_v \in \mathbb{R}^{F_{in}\times d_v^h}$ are learnable matrices mapping the input into different subspaces. The logit which indicates the attention strength of pixel i on pixel j is computed as:

$$l_{i,j} = \frac{q_i^T}{\sqrt{d_k^h}}\left(k_j + r^W_{j_x-i_x} + r^H_{j_y-i_y}\right), \tag{2}$$

where q_i is the *query* vector for pixel i, k_j is the *key* vector for pixel j, while $r^W_{j_x-i_x}$ and $r^H_{j_y-i_y}$ are learned embeddings for relative width $j_x - i_x$ and relative height $j_y - i_y$, respectively. Single head attention for head h is formulated as follows:

$$A_h = Softmax\left(\frac{QK^T + S_H^{rel} + S_W^{rel}}{\sqrt{d_k^h}}\right)V, \tag{3}$$

where $S_H^{rel}, S_W^{rel} \in \mathbb{R}^{HW \times HW}$ are matrices of learned positional encoding along height and width dimensions that satisfy $S_H^{rel}(i,j) = q_i^T r_{j_y - i_y}^H$ and $S_W^{rel}(i,j) = q_i^T r_{j_x - i_x}^W$. Thus, the above attention mechanism is able to attend feature sub-spaces and spatial positions simultaneously. The outputs A_i, from different heads $i, i \in [1,h]$ are concatenated together with the learnable matrix $W^o \in \mathbb{R}^{d_v \times d_v}$ to form multi-head attention mechanism, where d_v refers to the depth of *value*:

$$MHA(X) = Concat[A_1, ..., A_h]W^O. \tag{4}$$

The output is reshaped into size of (H,W,d_v) and concatenated with the output of the original convolution, forming the attention augmented convolution:

$$AAConv(X) = Concat[Conv(X), MHA(X)]. \tag{5}$$

3.2. LungAttn Neural Network

As illustrated in Fig. 5, the 3-channel spectrogram is fed into the proposed LungAttn neural network architecture for respiratory sound classification. A 7 × 7 convolution followed by max pooling is first applied to the original spectrogram to increase the number of channels, obtaining the primary feature maps. The ResNet-I layers reduce the size of the spectrogram while enlarging the receptive field. Residual connection is adopted to avoid the gradient vanishing problem (He et al. 2016). Subsequently, several cascade ResNet-II layers are used to learn the characteristics of lung sounds through time and frequency domains in the view of the STFT spectrogram. An AAResNet is inserted between ResNet-II layers. In the AAResNet, the first convolution layer is replaced by augmented attention convolution, which helps the neural network to focus on more relevant parts on the global scale rather than the local scale of the spectrogram. Lastly, we use group normalization (GN) (Wu and He 2018), rectified linear unit (ReLU) (Goodfellow, Bengio, and Courville 2016), global average pooling (GAP) (ibid.) and two fully connected layers to classify each respiratory sound into four types. Dropout is applied to fully connected layers to eliminate the overfitting problem.

As stated in Section 2.4, after mixup augmentation, some labels are not in one-hot format. Therefore, we propose the following loss function to improve the learning ability based on the distance between the labeled vector and predicted vector:

$$loss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{4}[p_c y_{nc} \cdot \log\sigma(x_{nc}) + (1 - y_{nc}) \cdot \log(1 - \sigma(x_{nc}))], \tag{6}$$

where x_{nc} and y_{nc} refer to the c_{th} element in predicted vector and label vector of the n_{th} cycle, respectively. $\sigma(\cdot)$ refers to sigmoid function, which is formulated as $\sigma(x) = 1/(1 + e^{-x})$. N is the total number of examples, and p_c is the weight of solely positive examples.

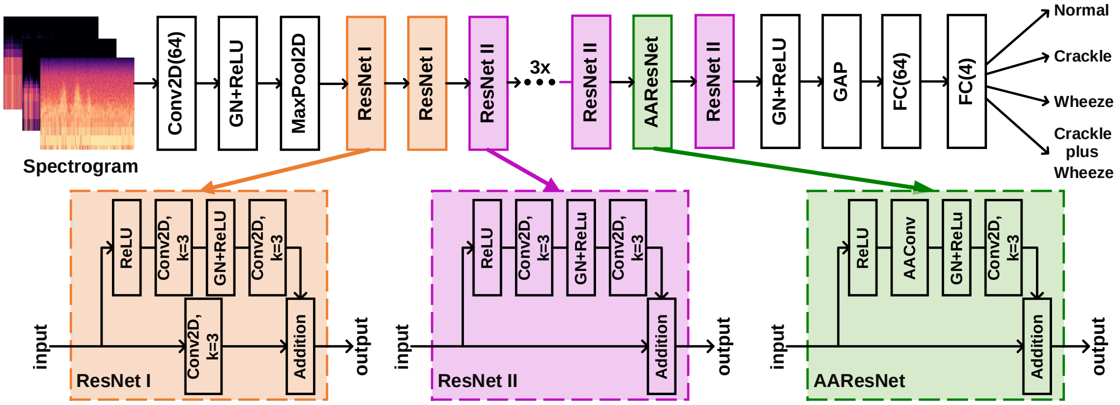


Figure 5. The proposed LungAttn neural network architecture.

4. Experimental Setup and Evaluation Methods

4.1. Experimental Setup:

We have implemented our LungAttn neural network model using Pytorch in Python3.6, and evaluated it on a 2.50GHz 12 cores Xeon Intel CPU based on a Linux machine with 128GB memory and an Nvidia GTX 2080TI graphics card. We initialize the trainable weights of the LungAttn model with Xavier initialization method (Glorot and Bengio 2010). Stochastic Gradient Descent (SGD) optimizer (Bottou 2010) is applied to train the model with a maximum mini-batch size of 16. The initial learning rate is set to 0.01, and the learning rate decay scheme decays exponentially for every 30 epochs. The dropout rate is 0.25 and the regularization coefficient is 1×10^{-4} to alleviate the overfitting problem. We have limited the maximum number of training epochs to 100 since the loss has reached a steady state.

We follow the official ICBHI 2017 dataset splitting method as well as the random 80/20 dataset splitting method to comprehensively validate our proposed framework. For the random 80/20 splitting method, there are no common patients seen in training set and testing set. Among the overall 126 patients in the ICBHI 2017 dataset, the training set involves recordings from 101 patients and the testing set involves recordings from the other 25 patients. A series of experiments is conducted to evaluate the effectiveness of our proposed techniques and compare our performance with the state-of-the-art works.

4.2. Evaluation Methods:

In this work, we have adopted the same evaluation method used in the official ICBHI 2017 contest (Rocha et al. 2019). The scoring method is defined as:

$$Sensitivity, S_e = \frac{P_c + P_w + P_b}{N_c + N_w + N_b}, \tag{7}$$

Table 2. Performance comparison for LungAttn using different high Q-factors based on validation set.

Q-factor	S_e	S_p	**Score**
2	51.73%	53.41%	52.57%
3	45.05%	**63.06%**	54.05%
4	**54.46%**	60.24%	**57.37%**
5	52.97%	60.00%	56.49%
6	52.24%	55.45%	53.84%

$$Specificity, S_p = \frac{P_n}{N_n}, \tag{8}$$

$$Score = \frac{S_e + S_p}{2}, \tag{9}$$

where P_c, P_w, P_b and P_n are the number of correctly predicted respiratory cycles in four types of lung sounds, respectively. N_c, N_w, N_b and N_n are the total number of instances in each type of lung sound cycles, respectively.

5. Experimental Results

5.1. Exploration of Q-factor

To improve the *Sensitivity* and *Specificity* of the model, there is a need to optimize the TQWT's parameters through the exploration of its Q-factor value. Notice that the oversampling rate r is set to 3 for both low and high Q-factor wavelet transform as recommended in (Selesnick 2011b; Bhattacharyya et al. 2017; Krishna et al. 2019). Since Q is required to be larger than 1 (Selesnick 2011b), for the low Q-factor wavelet, Q is set to 1. For the high Q-factor wavelet, we conducted experiments to evaluate the model performance with different high Q-factors. Since the exploration of Q-factor is a process of parameter estimation, we have divided the original training set into a new training set and validation set with the proportion of 80% and 20%, respectively. Considering the lack of training set, the new training and validation dataset division process is not based on patients. The newly obtained training set is used to train the LungAttn model, while the validation set is used to validate the model performance across different Q-factors. The experimental results based on the validation set are shown in Table 2. From Table 2, we observe that a high Q-factor of 4 performs optimally. With Q-factor increasing from 2 to 4, the *Score* increases from 52.57% to 57.37%. However, with Q-factor continuing to increase, the *Sensitivity* and *Specificity* start to decline, resulting the *Score* to decrease too. Thus, we set Q as 4 for the high Q-factor wavelet and conduct following experiments on independent test set.

Table 3. Performance comparison of different feature extraction methods with official ICBHI 2017 dataset splitting method.

Method	S_e	S_p	**Score**
Wavelet Transform	20.52%	63.00%	41.76%
Single STFT	**42.48%**	56.05%	49.26%
Dual TQWT and Single STFT	35.51%	68.12%	51.81%
Dual TQWT and Triple STFT	36.36%	**71.44%**	**53.90%**

5.2. Effectiveness of Feature Extraction

We performed several experiments to demonstrate the effectiveness of our proposed feature extraction method. For the first experiment, we only adopt wavelet transform on respiratory sounds. For the second experiment, we perform single STFT on respiratory sounds. For the third experiment, we perform dual TQWT with single STFT. That is, after decomposing the original signal into three signals using dual TQWT, we perform STFT with the same parameter settings for the three signals, obtaining three-channel spectrogram as the input of LungAttn model. It is worth noting that for the second and third experiments, Hanning window is used for single STFT with 20 ms window size and 50% overlap between adjacent windows. The experimental results are tabulated in Table 3. Single STFT method outperforms the wavelet transform method with 7.53% improvement on *Score*. Compared to single STFT feature extraction, dual TQWT with single STFT method improves S_p from 56.05% to 68.12% and *Score* from 49.26% to 51.81%. Using our proposed dual TQWT and triple STFT, S_p and *Score* are further improved to 71.44% and 53.90%, respectively.

5.3. Effectiveness of Mixup Data Augmentation

To demonstrate the effectiveness of our proposed mixup data augmentation method, we have conducted experiments to compare our methods with (i) the method without data augmentation (Experiment #1), and (ii) the traditional data augmentation methods (Experiment #2). For Experiment #1, we simply use the original imbalanced ICBHI 2017 dataset. For Experiment #2, we randomly generate a new dataset using crop (crop audio's segment randomly), loudness (adjust audio's volume), noise (inject noise), mask (mask audio's segment), pitch (adjust audio's pitch), and speed (adjust audio's speed) methods to generate 2,000 samples for each lung sound type. Table 4 reports the S_e, S_p and *Score* of our proposed methods using the datasets in Experiment #1 and Experiment #2. Clearly, traditional data augmentation methods improve the *Score* from 49.12% to 50.66% compared with the original imbalanced dataset. However, our proposed mixup data augmentation method has further improved the *Score* to 53.90%, demonstrating the effectiveness of our proposed innovation.

Table 4. Performance evaluation of different data augmentation methods with the official ICBHI 2017 dataset splitting method.

Data augmentation	S_e	S_p	Score
None (Exp.1)	34.66%	63.58%	49.12%
Traditional method (Exp.2)	35.26%	66.05%	50.66%
Mixup method	**36.36%**	**71.44%**	**53.90%**

Table 5. Performance comparison of different CNN models with the official ICBHI 2017 dataset splitting method .

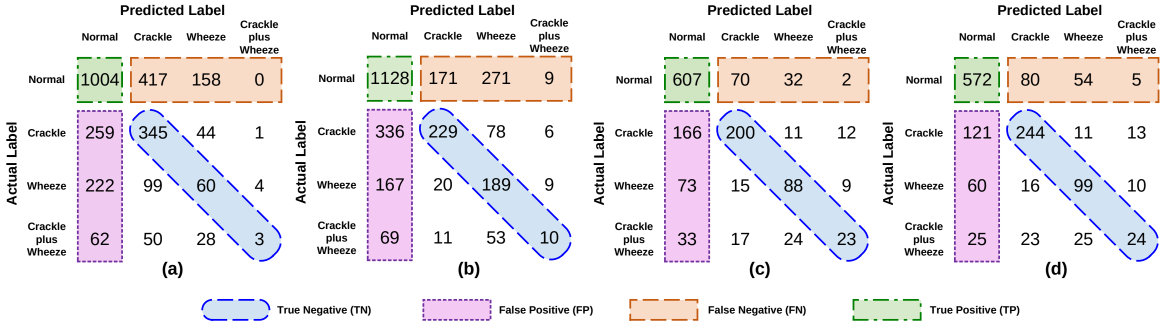
Model	S_e	S_p	Score
ResNet18	29.65%	70.68%	50.16%
VGG16	27.87%	74.54%	51.20%
MobileNet v2	25.49%	**75.81%**	50.65%
CNN w/o AA	33.64%	68.08%	50.86%
CNN with NL	34.75%	68.14%	51.45%
LungAttn	**36.36%**	71.44%	**53.90%**

5.4. Effectiveness of proposed LungAttn model

We have conducted experiments to compare the performance of our proposed LungAttn model and other popular CNN models, such as ResNet18, VGG16, MobileNet v2. To prove the effectiveness of our introduced augmented attention convolution, we also report the results of conventional ResNet and ResNet with non-local block (Ma, Xu, and Li 2020). The results are tabulated in Table 5. The ResNet18 model, VGG16 model and MobileNet v2 model achieved the *Score* of 50.16%, 51.20% and 50.65%, respectively. Our proposed LungAttn model has outperformed these popular CNN models with the best *Score* of 53.90%. The conventional ResNet without augmented attention convolution obtained the *Score* of 50.86%. After introducing Non-Local block in the ResNet, the *Score* is improved by 0.59%. Replacing the Non-Local block with augmented attention convolution, the *Score* is further enhanced from 51.45% to 53.90%.

5.5. Classification Performance and Comparison with the State-of-the-art Works

Fig. 6 illustrates the confusion matrices describing the prediction accuracy of each respiratory sound type on the testing datasets based on the official ICBHI 2017 dataset splitting method and random 80/20 dataset splitting method. The diagonal values show the number of correctly classified lung sound cycles with the predicted label and actual labels as the x-axis and y-axis, respectively. It can be seen from the confusion matrices that the best prediction performance occurs at "Normal" class while the worst performance has been observed for "Crackle plus Wheeze" class. With the proposed mixup data augmentation method, the classification accuracy of "Wheeze" and "Crackle



(a)

Actual Label \ Predicted Label	Normal	Crackle	Wheeze	Crackle plus Wheeze
Normal	1004	417	158	0
Crackle	259	345	44	1
Wheeze	222	99	60	4
Crackle plus Wheeze	62	50	28	3

(b)

Actual Label \ Predicted Label	Normal	Crackle	Wheeze	Crackle plus Wheeze
Normal	1128	171	271	9
Crackle	336	229	78	6
Wheeze	167	20	189	9
Crackle plus Wheeze	69	11	53	10

(c)

Actual Label \ Predicted Label	Normal	Crackle	Wheeze	Crackle plus Wheeze
Normal	607	70	32	2
Crackle	166	200	11	12
Wheeze	73	15	88	9
Crackle plus Wheeze	33	17	24	23

(d)

Actual Label \ Predicted Label	Normal	Crackle	Wheeze	Crackle plus Wheeze
Normal	572	80	54	5
Crackle	121	244	11	13
Wheeze	60	16	99	10
Crackle plus Wheeze	25	23	25	24

Figure 6. The confusion matrices of proposed LungAttn model based on (a) official ICBHI 2017 dataset splitting method without mixup (b) official ICBHI 2017 dataset splitting method with mixup (c) random 80/20 dataset splitting method without mixup, and (d) random 80/20 dataset splitting method with mixup.

plus Wheeze" has been improved significantly. For the official ICBHI 2017 dataset splitting method, the classification accuracy of "Wheeze" and "Crackle plus Wheeze" has been improved by 33.5% and 4.9%, respectively, while for random 80/20 dataset splitting method, the accuracy of "Crackle" and "Wheeze" has been improved by 11.3% and 5.9%, respectively.

As shown in Table 6, we compare our results against the state-of-the-art works (Jakovljević and Lončar-Turukalo 2018; Chambres, Hanna, and Desainte 2018; Ma et al. 2019; Ma, Xu, and Li 2020; Pham et al. 2021). For a comprehensive comparison, we reported the results of 4-class classification ("Normal", "Crackle", "Wheeze", and "Crackle plus Wheeze" classes), 2-class classification ("Normal" and "AbNormal" classes) tasks using official ICBHI 2017 dataset splitting method, and 4-class classification task using random 80/20 dataset splitting method. Column S_e, S_p and *Score* refer to the sensitivity, specificity and score reported in the state-of-the-art models (Jakovljević and Lončar-Turukalo 2018; Chambres, Hanna, and Desainte 2018; Serbes, Ulukaya, and Kahya 2018; Ma et al. 2019; Ma, Xu, and Li 2020; Pham et al. 2021) and our evaluation results from our LungAttn model. LungAttn model achieves S_e, S_p and *Score* of 36.36%, 71.44% and 53.90%, respectively on the 4-class classification task using official ICBHI 2017 dataset splitting method, outperforming all the state-of-the-art works on *Score* with an improvement of 1.64%. Besides, for 2-class classification task using the official ICBHI 2017 dataset splitting method, LungAttn model obtains S_e of 51.40%, S_p of 71.44% and *Score* of 61.42%, improving *Score* by 2.97% compared to the state-of-the-art works. For 4-class classification tasks using the random 80/20 dataset splitting method, LungAttn model reports S_e, S_p and *Score* of 54.69%, 80.45% and 67.57%, respectively, outperforming the state-of-the-art works.

5.6. Discussion

We propose a new method that relies on characterization of different adventitious lung sounds to extract features, including decomposition of the original sound and

Table 6. Performance Comparison with state-of-the-art models.

	Reference	S_e	S_p	**Score**
Official, 4-class	(Jakovljević and Lončar-Turukalo 2018)	-	-	39.56%
	(Chambres, Hanna, and Desainte 2018)	20.81%	**78.05%**	49.43%
	(Serbes, Ulukaya, and Kahya 2018)	-	-	49.86%
	(Ma et al. 2019)	31.12%	69.20%	50.16%
	(Ma, Xu, and Li 2020)	**41.32%**	63.20%	52.26%
	(Pham et al. 2021)	26.00%	68.00%	47.00%
	Ours	36.36%	71.44%	**53.90%**
Official, 2-class	(Chambres, Hanna, and Desainte 2018)	33.15%	**78.05%**	55.60%
	(Ma et al. 2019)	47.70%	69.20%	58.45%
	(Pham et al. 2021)	-	-	54.10%
	Ours	**51.40%**	71.44%	**61.42%**
Random, 4-class	(Kochetov et al. 2018)	58.43%	73.00%	65.70%
	(Acharya and Basu 2020)	48.63%	**84.14%**	66.38%
	(Ma, Xu, and Li 2020)	**63.69%**	64.73%	64.21%
	Ours	54.69%	80.45%	**67.57%**

time-frequency representation conversion. The experimental results show our method effectively captures features for different lung sounds. Furthermore, the mixup data augmentation is a simple and efficient way to augment adventitious lung sound samples, which can be potentially used in other fields like heart sound classification. Attention mechanism introduced in our network outperforms vanilla ResNet, indicating the feasibility of attention mechanism in biomedical signal classification. The evaluation results on ICBHI 2017 dataset shows classification performance improvement, which is supposed to be greater with future larger respiratory sound dataset. There are some limitations of our approach. Currently, we are working on building a new respiratory dataset with high-quality lung sounds. In the future work, we expect to use the new dataset combined with previous publicly available datasets to validate our method and gradually achieve clinical application. Furthermore, we will explore a method for self-adaptive Q-factor instead of finding the optimal Q-factor by experiments in the future.

6. Conclusions

This paper presents a dual TQWT with triple STFT to transform the lung sounds into a multi-channel spectrogram, which is fed into our proposed adventitious lung sound classification model, LungAttn. To capture the long spatial and temporal interval relationship in the spectrogram, we introduce attention mechanism to our model by combining augmented attention convolution with ResNet blocks. The mixup data augmentation method is used to address the imbalance problem in the ICBHI 2017 dataset. Based on the official ICBHI 2017 evaluation method, our experimental results have achieved the S_e, S_p and *Score* of 36.36%, 71.44% and 53.90%, respectively, with 1.69% - 14.34% improvement on *Score* compared with the state-of-the-art works.

7. Acknowledgements

This research is supported in part by the National Key Research and Development Program of China under Grant No. 2019YFB2204500 and in part by the Science, Technology and Innovation Action Plan of Shanghai Municipality, China under Grant No. 1914220370.

References

Acharya, J. and A. Basu (2020). "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning". In: *IEEE Transactions on Biomedical Circuits and Systems*, pp. 1–1.

Aziz, S. et al. (2019). "An Automated System towards Diagnosis of Pneumonia using Pulmonary Auscultations". In: *International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, pp. 1–7.

Bello, I. et al. (2019). "Attention augmented convolutional networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3286–3295.

Bhattacharyya, A. et al. (2017). "Tunable-Q wavelet transform based multiscale entropy measure for automated classification of epileptic EEG signals". In: *Applied Sciences* 7.4, p. 385.

Bobin, Jé. et al. (2007). "Morphological component analysis: An adaptive thresholding strategy". In: *IEEE transactions on image processing* 16.11, pp. 2675–2681.

Bohadana, A., G. Izbicki, and S. Kraman (2014). "Fundamentals of lung auscultation". In: *New England Journal of Medicine* 370.8, pp. 744–751.

Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent". In: *Proceedings of COMPSTAT'2010*. Springer, pp. 177–186.

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

Chen, H. et al. (2019). "Triple-Classification of Respiratory Sounds Using Optimized S-Transform and Deep Residual Networks". In: *IEEE Access* 7, pp. 32845–32852.

Cukic, V. et al. (2012). "Asthma and chronic obstructive pulmonary disease (COPD)–differences and similarities". In: *Materia socio-medica* 24.2, p. 100.

Demir, F., A. M. Ismael, and A. Sengur (2020). "Classification of Lung Sounds With CNN Model Using Parallel Pooling Structure". In: *IEEE Access* 8, pp. 105376–105383.

García-Ordás, M. T. et al. (2020). "Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data". In: *Sensors* 20.4, p. 1214.

Glorot, X. and Y. Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Journal of Machine Learning Research* 9, pp. 249–256.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning.* MIT press.

He, K. et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Jakovljević, N. and T. Lončar-Turukalo (2018). "Hidden markov model based respiratory sound classification". In: *Precision Medicine Powered by pHealth and Connected Health*, pp. 39–43.

Ko, T. et al. (2015). "Audio augmentation for speech recognition". In: *Annual Conference of the International Speech Communication Association*, pp. 1–4.

Kochetov, K. et al. (2018). "Noise masking recurrent neural network for respiratory sound classification". In: *International Conference on Artificial Neural Networks*, pp. 208–217.

Krishna, A. H. et al. (2019). "Emotion classification using EEG signals based on tunable-Q wavelet transform". In: *IET Science, Measurement & Technology* 13.3, pp. 375–380.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25, pp. 1097–1105.

Lia, C. P. et al. (2016). "Computerized wheeze detection in young infants: comparison of signals from tracheal and chest wall sensors". In: *Physiological measurement* 37.12, pp. 2170–2180.

Liu, X. et al. (2015). "Detection of adventitious lung sounds using entropy features and a 2-D threshold setting". In: *International Conference on Information, Communications and Signal Processing (ICICS)*, pp. 1–5.

Lozano, M., J. A. Fiz, and R. Jané (2016). "Automatic differentiation of normal and continuous adventitious respiratory sounds using ensemble empirical mode decomposition and instantaneous frequency". In: *IEEE Journal of Biomedical and Health Informatics* 20.2, pp. 486–497.

Luis, P. and W. Jason (2017). "The effectiveness of data augmentation in image classification using deep learning". In: *Computer Vision and Pattern Recognition*, pp. 1–8.

Ma, Y., X. Xu, and Y. Li (2020). "LungRN+NL: An improved adventitious lung sound classification using non-local block ResNet neural network with mixup data augmentation". In: *Proc. Interspeech*, pp. 2902–2906.

Ma, Y. et al. (2019). "Live Demo: LungSys - Automatic Digital Stethoscope System For Adventitious Respiratory Sound Detection". In: *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–1.

Ma, Y. et al. (2019). "LungBRN: A smart digital stethoscope for detecting respiratory disease using bi-ResNet deep learning algorithm". In: *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4.

Marques, A. and B. Anne (2009). “The reliability of lung crackle characteristics in cystic fibrosis and bronchiectasis patients in a clinical setting”. In: *Physiological measurement* 30.9, pp. 903–912.

Messner, E. et al. (2020). “Multi-channel lung sound classification with convolutional recurrent neural networks”. In: *Computers in Biology and Medicine* 122, p. 103831.

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

Pham, L. D. et al. (2021). “CNN-MoE based framework for classification of respiratory anomalies and lung disease detection”. In: *IEEE Journal of Biomedical and Health Informatics.*

Piirila, P. and A. R. Sovijarvi (1995). “Crackles: recording, analysis and clinical significance”. In: *European Respiratory Journal* 8.12, pp. 2139–2148.

Pramono, R. X. A., S. A. Imtiaz, and E. Rodriguez-Villegas (2019). “Evaluation of features for classification of wheezes and normal respiratory sounds”. In: *PloS one* 14.3, e0213659.

Rabinovich, A. et al. (2007). “Objects in context”. In: *IEEE International Conference on Computer Vision*, pp. 1–8.

Rocha, B. M. et al. (2019). “An open access database for the evaluation of respiratory sound classification algorithms”. In: *Physiological measurement* 40.3, pp. 1–28.

Rocha, V., M. Cristina, and M. Alda (2016). “Computerized respiratory sound analysis in people with dementia: a first-step towards diagnosis and monitoring of respiratory conditions”. In: *Physiological measurement* 37.11, pp. 2019–2092.

Rollins, R. et al. (2015). “Discrete conditional phase-type model utilising a multiclass support vector machine for the prediction of retinopathy of prematurity”. In: *IEEE International Symposium on Computer-Based Medical Systems*, pp. 250–255.

Selesnick, I. W. (2011a). “Sparse signal representations using the tunable Q-factor wavelet transform”. In: *Wavelets and Sparsity XIV*. Vol. 8138. International Society for Optics and Photonics, 81381U1–81381U15.

– (2011b). “Wavelet transform With tunable Q-factor”. In: *IEEE Transactions on Signal Processing* 59.8, pp. 3560–3575.

Sengupta, N., M. Sahidullah, and G. Saha (2016). “Lung sound classification using cepstral-based statistical features”. In: *Computers in Biology and Medicine* 75, pp. 118 –129.

Serbes, G., S. Ulukaya, and Y. Kahya (2018). “An automated lung sound preprocessing and classification system based onspectral analysis methods”. In: *Precision Medicine Powered by pHealth and Connected Health*, pp. 45–49.

Shivnarayan, P. and P. Ram (2013). “Constrained tunable-Q wavelet transform based analysis of cardiac sound signals”. In: *AASRI Procedia* 4, pp. 57–63.

Shivnarayan, P. and P. Trilochan (2017). “Detection of epileptic seizure using Kraskov entropy applied on tunable-Q wavelet transform of EEG signals”. In: *Biomedical Signal Processing and Control* 34, pp. 74–80.

Shuvo, S. B. et al. (2020). “A lightweight cnn model for detecting respiratory diseases from lung auscultation sounds using emd-cwt-based hybrid scalogram”. In: *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1.

Vaswani, A. et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems*, pp. 5998–6008.

Wiśniewski, M. and T. P. Zieliński (2015). “Joint application of audio spectral envelope and tonality index in an E-asthma monitoring system”. In: *IEEE Journal of Biomedical and Health Informatics* 19.3, pp. 1009–1018.

Wu, Y. and K. He (2018). “Group normalization”. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19.

Xie, J. et al. (2020). “Critical care crisis and some recommendations during the COVID-19 epidemic in China”. In: *Intensive care medicine*, pp. 1–4.

Zhang, H. and M. Cisse (2018). “Mixup: beyond empirical risk minimization”. In: *International Conference on Learning Representations*, pp. 1–13.

Zhang, K. et al. (2015). “The detection of crackles based on mathematical morphology in spectrogram analysis”. In: *Technology and Health Care* 23.s2, S489–S494.