IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 10, 2019, accepted September 19, 2019, date of publication September 26, 2019,
date of current version October 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2943492

Lung Sound Recognition Algorithm Based on VGGish-BiGRU

LUKUI SHI[1,2], KANG DU[1], CHAOZONG ZHANG[3], HONGQI MA[1], AND WENJIE YAN[1,2]
[1]School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China
[2]Key Laboratory of Big Data Computing of Hebei, Tianjin 300401, China
[3]Hebei Institute of Scientific and Technical Information, Shijiazhuang 050000, China

Corresponding author: Wenjie Yan (wenjieyanhit@163.com)

This work was supported in part by the Key Project of the Joint Fund of the National Natural Science Foundation of China under Grant U1813222, in part by the National Key Research and Development Program of China under Grant 2018YFB1306900, in part by the NSF of Hebei Province of China under Grant F2017202145, in part by the TSTC of Tianjin of China under Grant 16JCYBJC15600, in part by the National Natural Science Foundation of China under Grant 61702157, in part by the Natural Science Foundation of Tianjin City of China under Grant 16JCQNJC00400, and in part by the Big Data Project of Hebei Institute of Scientific and Technical Information.

ABSTRACT Pulmonary breathing sound plays a key role in the prevention and diagnosis of the lung diseases. Its correlation with pathology and physiology has become an important research topic in the pulmonary acoustics and the clinical medicine. However, it is difficult to fully describe lung sound information with the traditional features because lung sounds are complex and nonstationary signals. And the traditional convolutional neural network cannot also extract the temporal features of the lung sounds. To solve the problem, a lung sound recognition algorithm based on VGGish-BiGRU is proposed on the basis of transfer learning, which combines VGGish network with the bidirectional gated recurrent unit neural network (BiGRU). In the proposed algorithm, VGGish network is pretrained using audio set, and the parameters are transferred to VGGish network layer of the target network. The temporal features of the lung sounds are extracted through retraining BiGRU network with the lung sound data. During retraining BiGRU network, the parameters in VGGish layers are frozen, and the parameters of BiGRU network are fine-tuned. The experimental results show that the proposed algorithm effectively improves the recognition accuracy of the lung sounds in contrast with the state-of-the-art algorithms, especially the recognition accuracy of asthma.

INDEX TERMS BiGRU, lung sound recognition, Mel spectrogram, transfer learning, VGGish.

I. INTRODUCTION

Pulmonary auscultation is one of the effective methods to diagnose the lung diseases. Early lesions of the lungs can be earlier detected by diagnosing the lung diseases with stethoscopes. However, the frequency range of the lung sounds is about 100Hz to 2000Hz, while the human ear is only sensitive to the frequency range of 1000Hz to 2000Hz. Therefore, it is easy for the artificial auscultation to lose the important information of the low frequency part of the lung sounds. And the auscultation results are affected by the doctor's medical experience, hearing condition, and external environment. It is possible for some lung diseases to have a risk of the misdiagnosis and missed diagnosis.

With the advancement of digital signal processing and artificial intelligence technologies, the traditional acoustic stethoscope has been gradually replaced with the electronic stethoscope, which provides an opportunity to solve the above problems. The electronic stethoscope can store the lung sound signals and transmit the lung sound signals to the computer when the doctor uses the electronic stethoscope to diagnose the lung diseases. The lung sound signals can be recognized by analyzing the time-frequency characteristics of the signals and building the recognition model. It can further predict the healthy status of the lung. Therefore, it can not only improve the diagnostic accuracy, but also improve the diagnostic efficiency by using the artificial intelligence technologies to analyze and identify the lung sound signals collected by the electronic stethoscope. It is greatly significant for the prevention and treatment of the lung diseases.

In the early stage of the lung sound recognition, the lung sounds were recognized by using the traditional machine learning methods. The crackles were effectively recognized

The associate editor coordinating the review of this manuscript and approving it for publication was Malik Jahan Khan.

through extracting the time-frequency distribution features of the peaks from the lung sounds after being executed Hilbert Huang transform in [1]. Fast Fourier transform was used to construct feature vectors and lung sounds are classified by the self-organizing map in [2]. Short-Time Fourier Transform was utilized to extract features from lung sounds and support vector machine (SVM) was used for wheeze recognition in [3]. Lung sound signals were decomposed into the frequency subbands using wavelet transform and a set of statistical features was extracted from the subbands to represent the distribution of wavelet coefficients. Lung sounds are classified into six categories by using artificial neural network in [4]. The wavelet analysis was used to extract the features of the lung sounds, and SVM was utilized to classify the lung sounds in [5]. Wavelet transforms and neural networks are used to classify asthmatic breath sounds in [6]. The wavelet coefficients were extracted from the lung sounds, and neural network and SVM were used to identify the lung sounds in [7]. The lung sounds are extracted wavelet coefficients and linear discriminant analysis to reduce the dimension of wavelet coefficients in [8]. The lung sounds are classified by BP neural network. The wavelet packet decomposition was used to get the energy of the lung sounds with different frequency range, which was taken as features to recognize four kind of lung sounds including the normal, tracheitis, pneumonia and asthma in [9].

The method was obtained based on the bispectrum, which extracted the spectral peak, the spectral peak interval and the slice energy in [10]. A set of features based on temporal characteristics of filtered narrowband signal were proposed to classify respiratory sounds into normal and continuous adventitious types in [11]. The lung sound signal was mapped onto a rich spectro-temporal feature space and was classified by using SVM in [12]. The method combining linear predictive cepstral coefficient (LPCC) with the wavelet decomposition was put forward in [13]. It could effectively recognize the polyphonic lung sounds and the sharp lung sounds. Mel frequency cepstral coefficients (MFCC) were extracted from the pre-processed pulmonary acoustic signals. And the performance of SVM and KNN classifiers in diagnosis respiratory pathologies was compared by using respiratory sounds from R.A.L.E database in [14]. MFCC features of the lung sounds to recognize the lung sounds in [15]. Results showed that the method outperformed commonly used wavelet-based features as well as standard cepstral coefficients including MFCCs. The features were extracted from the wavelet coefficients by MFCC and non-Gaussian power, which was used to detect cough and burst sounds in the lung sounds and was taken as a basis for judging children's pneumonia in [16].

Gaussian mixture model was established to distinguish the normal lung sounds and the abnormal lung sounds in [17]. Multi-layer perceptron was employed to classify the lung sound signals in [18].The genetic BP neural network was proposed to recognize the lung sounds in [19]. And experiments showed that the improved genetic BP neural network was prior to the traditional BP neural network. The linear parameterized method of multi-channel lung sound information was employed to recognize the lung sound signals in [20]. The neural network was used during the classification process. And better recognition results were achieved. A set of statistical features was computed from each subband of lung sounds and applied to ANN and SVM classifiers to classify normal and asthmatic subjects on 4-channel data in [21]. Different parameterization techniques for lung sounds acquired on the whole posterior thoracic surface for normal versus abnormal lung sound classification were assessed in [22]. Some methods including logistic regression, decision tree, K near-neighbor, SVM, and naive Bayesian were compared and analyzed in the application of the lung sound recognition in [23].

In recent years, it provides a new identification method for the medical lung sound diagnosis technology with the development of deep learning. A lung sound recognition method based on convolutional neural network (CNN) was proposed in [24]. The method extracted MFCC features, and a two-layer convolutional neural network (2L-CNN) was used to train and recognize the lung sounds. Experimental results showed that the recognition method based on CNN is prior to the method based on SVM. MFCC features were utilized to identify the lung sounds by a five-layer convolutional neural network (5L-CNN), and better recognition results were obtained in [25]. Short time Fourier transform was used to analyze time-frequency features of the lung sounds and the lung sounds were classified into three categories by combining two-layer convolutional neural networks with two-layer full connections in [26]. A combined model framework DNN-HMM was proposed to identify the normal and abnormal lung sounds, which combined deep neural network with the hidden Markov model in [27]. The model could greatly improve the classification performance. A method for the identification of wheeze, crackle, and normal sounds was proposed, which uses the optimized S-transform and deep residual networks in [28]. Better results were obtained.

Although these methods can better recognize the lung sounds, most of them are built based on the artificial features or small datasets because there is not enough big open lung sound dataset. It leads to high dependence on data and features, which causes weak generalization ability. At the same time, the lung sound signal as a kind of time series signal contains rich temporal characteristics, which is an important feature of the lung sounds. However, the temporal feature is not fully reflected in the above methods.

In this paper, a lung sound recognition algorithm based on deep learning and transfer learning is proposed. In the algorithm, VGGish is firstly used to overcome the dependence of the algorithms on data and features. Secondly, the temporal feature of the lung sound signals is captured by taking the bidirectional gated recurrent unit neural network (BiGRU) as the retraining layer of transfer learning. Experimental results in section IV also show that the recognition accuracy is greatly improved.

II. LUNG SOUND RECOGNITON MODEL BASED ON VGGISH-BIGRU

A. MODEL ARCHITECTURE

With the rapid development of artificial intelligence, deep neural networks have achieved great success in many fields. CNN and recurrent neural network (RNN) have been widely applied in audio recognition. However, it is easy to generate over-fitting directly using the depth model due to the insufficient lung sound data. Therefore, it is a better choice to use transfer learning to improve the generalization ability of the model. VGGish network can better handle audio data, and RNN can better process time series data. From the view of the spatial domain and the time domain, a lung sound recognition model based on VGGish-BiGRU is proposed by combining VGGish convolutional neural network with BiGRU recurrent neural network. It is shown in Fig. 1.

In the model, the source domain data of transfer learning is Audio Set [29], which is a large-scale labeled audio dataset that Google opened in 2017. And the target domain data is the self-collected lung sound data. The model mainly

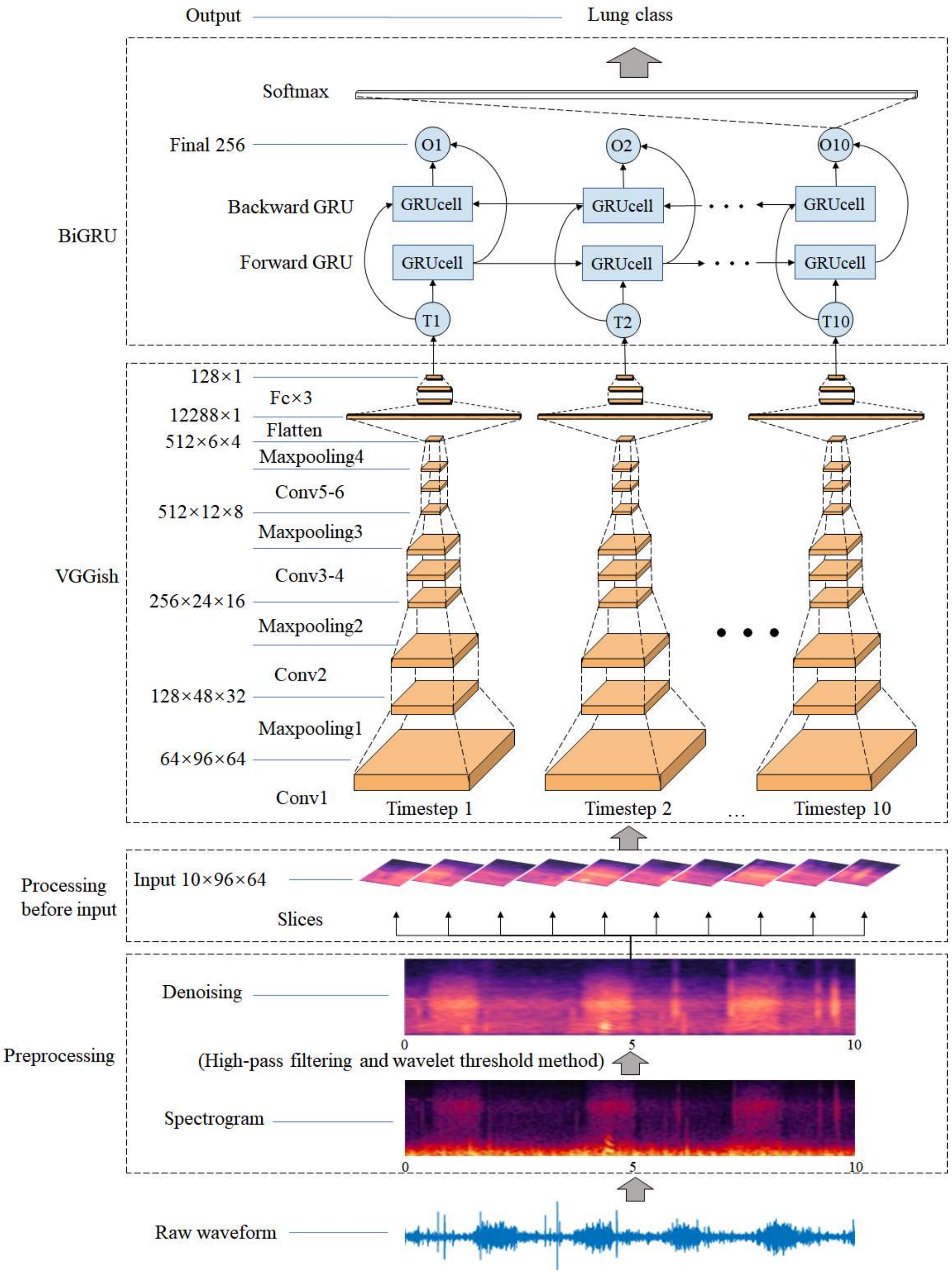


FIGURE 1. The lung sound recognition model based on VGGish-BiGRU.

includes two parts: VGGish convolution neural network and bidirectional GRU network. VGGish network part is transferred from VGGish network in the source domain, and the parameters of the corresponding trained network are loaded. The network outputs a 128-dimensional feature vector for each time series information. The bidirectional GRU network part inputs the features extracted by the transfer layer into BiGRU network and locally fine-tunes BiGRU layer. Here, the number of the hidden layer neurons in BiGRU is 128 × 2.

Since the similarity between the target data and the source domain data is lower, it is very indispensable to retrain the model with the target data. To this end, BiGRU network is taken as the retraining layer in the model, and the problem of insufficient data is compensated by freezing the parameters of the network layer of the pretraining model. RNN is chosen as the retraining part of the model since RNN has a strong ability to capture the time series features of the signals, and can pay more attention to the data context. As a kind of time series data, the temporal relationship existing in the lung sounds can effectively be captured through retraining BiGRU. The process of the model is as follows:

1) Train VGGish network on Audio Set and save the training parameters of the model. Since Audio Set is too large, we directly utilize the .ckpt model file, which has been trained by Google on VGGish.

2) Transfer the model parameters trained on the source domain to the target model.

3) Retrain VGGish-BiGRU model on the target domain. During retraining the model, the parameters of VGGish model are frozen, and the parameters in BiGRU model are only fine-tuned.

B. PREPROCESSING OF THE LUNG SOUNDS

During the process of detecting the lung sounds, it is unavoidable to mix the low frequency noise such as noise of the collecting devices, friction sound of the internal organs of the human body and so on. And the frequency range is more concentrated. It can be considered that there are no lung sound signals under 100Hz since the frequency band of the lung sound signals is 100HZ to 2000Hz. Therefore, the low frequency noise under100Hz can be removed by high-pass filters. At the same time, the lung sounds also mix a lot of the heart sounds besides the low frequency noise. The frequency band of the heart sounds in the lung sounds is 5HZ to 600Hz, which highly coincides with the low frequency part of the lung sounds. It is difficult to remove the interference of the heart sounds under without damaging the lung sounds by simple filtering. To remove the noise of the lung sounds, a hybrid de-noising technique is used. At first, the low frequency noise is deleted by a fourth-order Butterworth high-pass filter [30]. Then the heart sounds are removed by the wavelet threshold method.

The wavelet threshold method uses the multi-scale analysis method to decompose the target signal and extract different wavelet coefficients. And the wavelet coefficients are partitioned from the level of the wavelet decomposition by setting the appropriate wavelet threshold. The wavelet coefficients below the threshold are cleared, and the wavelet coefficients above the threshold are retained. Finally, the pure lung sound signals can be obtained by reconstructing the wavelet coefficients.

The results by removing the noise from the lung sounds with the above method are shown in Fig. 2 and Fig. 3.

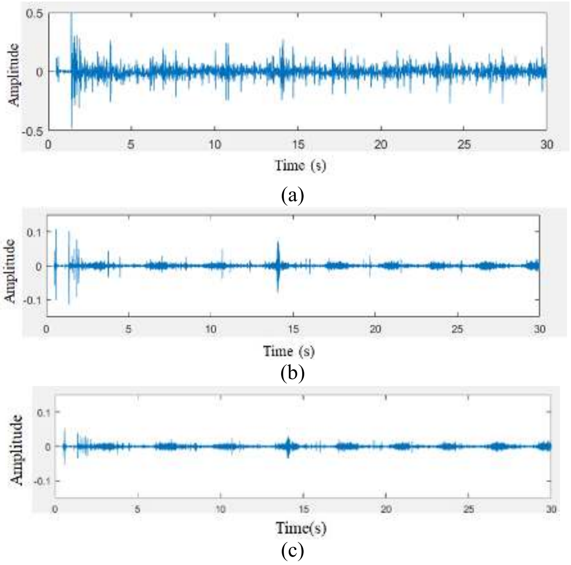


FIGURE 2. Comparison of the lung sound signals before and after Denoising (a) Original lung sound signal (b) The lung sound signal after high-pass filtering (c) The lung sound signal removed the heart sounds.

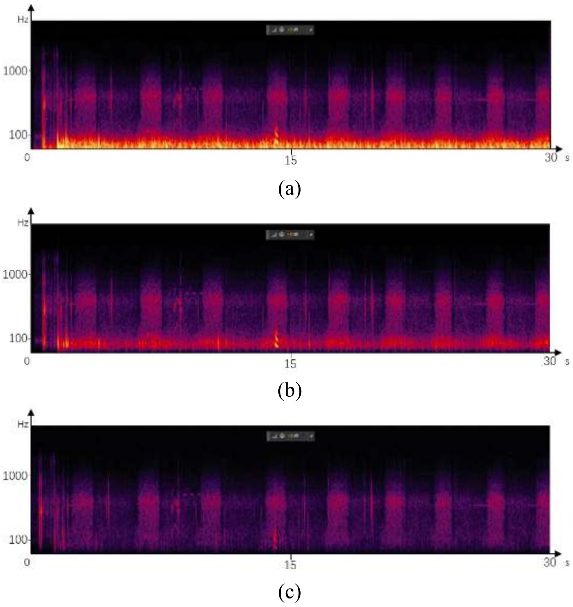


FIGURE 3. Changes in the spectrogram of the lung sound signals (a) The spectrogram of the original lung sound signal (b) The spectrogram of the lung sound signal by high-pass filtering (c) The spectrogram of the lung sound after deleting the heart sounds.

As shown in the above figures, the low frequency region of the spectrogram of the lung sounds mainly includes the ambient sounds and other low frequency noise. Therefore, it is necessary to remove the low frequency noise by the high-pass filtering. Simultaneously, it can be found from the spectrogram after the high-pass filtering that there are still a lot of heart sound components in the lung sound signals. And the energy of the heart sound signals is generally higher than that of the lung sound signals. If the heart sounds are not removed, the heart sounds with higher energy will greatly affect the recognition effect of the relatively weak lung sound signals. It can be seen from the results that the low frequency noise and the heart sound components are effectively deleted.

C. INPUT PROCESSING OF THE LUNG SOUND DATA

The original lung sound signals cannot be directly used as the input of the model. It is necessary to perform time frequency analysis on the lung sounds to obtain their input features. The model takes Mel spectrogram features of the lung sounds as the input of the network. To extract the Mel spectrogram features, the following processes need to be completed.

Firstly, pre-emphasis, framing and windowing are performed on the lung sounds, where the window function is a Hamming window, the window size is 25ms, and the step length is 10ms. The lung sounds are transformed from the time domain to the frequency domain by performing STFT. In the time-frequency transformation, it is necessary to define a range of the frequency domain of the lung sounds to obtain the main frequency domain information. The sampling frequency of the lung sounds is 4000Hz, so Nyquist frequency of the signal is 2000Hz. From the spectrogram of the lung sounds, the frequency components of the lung sound signals above 1000 Hz are few. And the starting frequency is 100Hz. Therefore, we set the frequency range of the lung sounds to be extracted the features to 100Hz to 1000Hz.

Secondly, the spectrum line energy of the lung sounds is filtered by using Mel filter bank $H_m(k)$. The function of filters is as follows:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1)-k}{f(m+1)-f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{1}$$

where $0 \le m \le M$, and M is the number of filters. Its center frequency $f(m)$ can be expressed as:

$$f(m) = \left(\frac{N}{f_s}\right) F_{mel}^{-1}\left(F_{mel}(f_l) + m\frac{F_{mel}(f_h) - F_{mel}(f_l)}{M+1}\right) \tag{2}$$

where f_l is the lowest frequency in the frequency domain of filters, and f_h is the highest frequency. N is the length of Fourier transform, and f_s is the sampling frequency. F_{mel} is Mel frequency. The transform formula between F_{mel} and the ordinary frequency f is as follows:

$$F_{\mathrm{mel}} = 2595 log(1 + f/700) \tag{3}$$

Then the inverse function F_{mel}^{-1} of F_{mel} is:

$$F_{mel}^{-1}(b) = 700(e^{b/2595} - 1) \tag{4}$$

where b is the real frequency.

The amplitude spectrum obtained by short time Fourier transform is separately multiplied with each filter. And all items are accumulated. Each frame contains 64 Mel bands while extracting Mel spectrogram features.

Finally, Mel spectrogram is achieved by taking the log value of the energy and expanding it in the time domain. Mel spectrogram of a lung sound signal is given in Fig. 4.

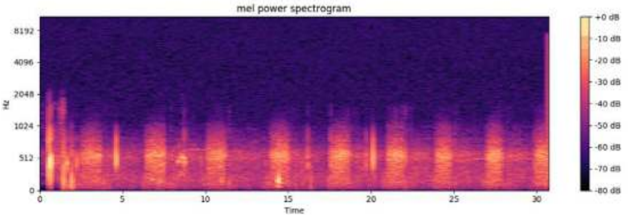


FIGURE 4. Mel spectrogram of the lung sound.

The length of the collected lung sound is about 30 seconds. To ensure the consistency of the input signal dimensions, the lung sound data need to be sliced before being input into the network. In the process, we need ensure that the length of each slice is the same, and there is at least one complete respiratory cycle in each slice. According to the respiratory characteristics of the human body, one cycle continues about 2-4s. Considering that the respiratory cycle of each person is slightly different, and the starting time of each audio is not always the starting time of one respiratory cycle, the length of each slice is set to 10s. Thus, all slices have the same length, and each slice exists 1 to 2 complete respiratory cycles at least.

At the same time, the retraining part of the model is BiGRU, which is based on the data time series relationship as the input. Therefore, 10s Mel spectrum is firstly segmented in the time domain when inputting data into the network. 10s lung sound is divided into 10 time series with 1s length. Each second audio is segmented into 96 frames and 64 Mel features are extracted per frame. Thus, the last input of the network is a 10 × 96 × 64 matrix. 10 represents ten time scales, 96 denotes the number of frames in each time scale, and 64 is the number of Mel features.

D. TRANSFERRING THE MODEL PARAMETERS FROM THE SOURCE DOMAIN TO THE TARGET DOMAIN

The model parameters can be transferred to the target network after the source domain model completes the training. The transferring model is shown in Fig. 5. VGGish network in the dotted box on the right needs to be transferred to the target model. In the transferred process, two networks must have the same structure to fully match their parameters. Therefore, VGGish network with the same structure is firstly built on the target network. Secondly, VGGish network model trained

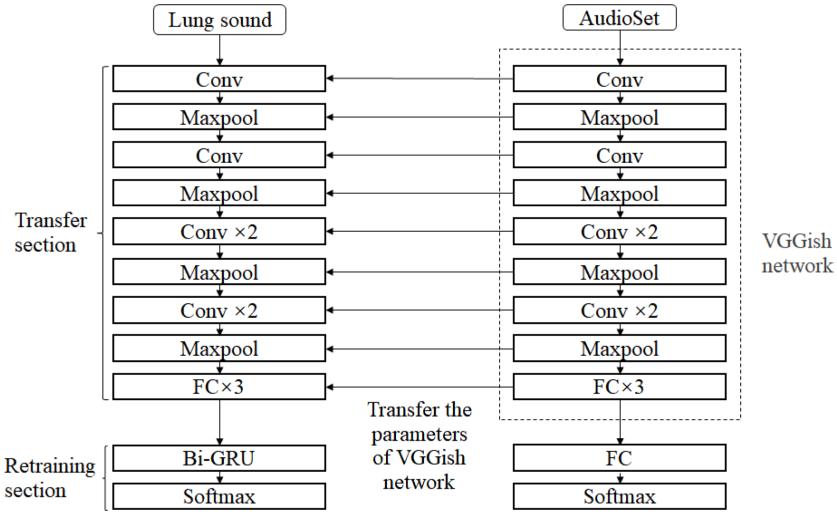


FIGURE 5. Transfer learning model based on VGGish-BiGRU.

and the reserved parameters are loaded by loading the tensor node of the checkpoint in TensorFlow. Thus, the parameters of the corresponding layers in the source domain network are transferred to the target network.

There are two main reasons for freezing the source domain model parameters after transferring the source domain model parameters to the target domain. Firstly, the insufficient target data can be compensated by freezing the network parameters of the pretrained model. Secondly, Audio set as a general audio dataset, contains a great number of samples and many categories. By transferring the model parameters, the target network has learned the basic knowledge, which is obtained by training the large scale audio data. The weights are randomly initialized while retraining the network. A large number of weight parameters will be propagated back through the network if the parameters of the transfer layer are not frozen, which will damage the feature representation learned in advance. In this way, the meaning of transfer learning is lost. According to the above reasons, the parameters of the transfer layers are frozen, and VGGish network is used as a feature extractor during the training model. VGGish network is only performed forward propagation and is not executed backpropagation. The retraining layer is trained with the target data.

III. RETRAINING OF BIGRU NETWORK

In the forward propagation of the network, VGGish network will output a 10 × 128 feature vector for each 10 second lung sound, and randomly initializes a bidirectional GRU network. The 10 × 128 feature vector will be input into BiGRU network in time series. The structure of BiGRU is given in Fig. 6.

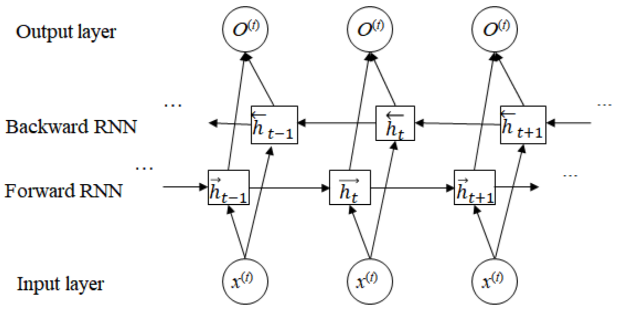


FIGURE 6. The structure of BiGRU.

As shown in the figure 6, BiGRU consists of a forward and a backward gated recurrent unit (GRU) neural network, and they are connected to the same output layer. GRU is an important variant of LSTM, which simplifies the structure of LSTM. In GRU, there are only two gates called an update gate and a reset gate. It makes the network parameters less and converge more easily. The structure of GRU is shown in Fig. 7.

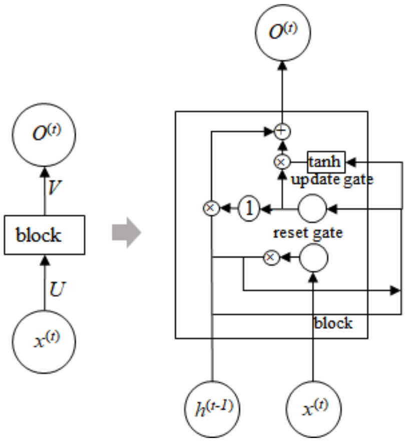


FIGURE 7. The structure of GRU.

In GRU, the update gate controls how much information from the previous hidden state will carry over to the current hidden state. The update gate is computed by

$$a_z^t = f(\sum w_{iz}[h^{t-1}, x_i^t]) \tag{5}$$

where x_i^t and h^{t-1} are respectively the input and the previous hidden state. w_{iz} is the weight between the input and the update gate. f(·) adopts the sigmoid function in [31]. The update gate in the GRU uses the sigmod function to help update and filter the data information, since any number multiplied by 0 is 0, which causes the value to disappear or be forgotten. Any number multiplied by 1 is the same value, so the value remains the same or keep. Through the sigmod function network, it can be understood which data is not important may be discarded or which data is important to be maintained. In the paper, we found that the better results could be gained by using the sigmoid function in experiments of section IV.

The reset gate mainly determines how much information will be forgotten, which represents the importance of the previous hidden state to the current hidden state. The reset gate is computed by

$$a_r^t = f(\sum w_{ir}[h^{t-1}, x_i^t]) \tag{6}$$

where w_{ir} is the weight between the input and the reset gate.

In the backpropagation of VGGish-BiGRU network, only BiGRU layers are fine-tuned since the parameters of VGGish layers are frozen. For the classification problem, the softmax loss is taken as the loss function of the backpropagation. The output of the softmax layer is computed by

$$S_i = \exp(x_i) / \sum_{k=1}^{T} \exp(x_k) \tag{7}$$

where S_i is the predicted probability of the sample belonging to the i^{th} class lung sound, x expresses the output vector of the previous layer, x_i is the i^{th} attribute of x, and T is the number of categories. Then the softmax cross entropy loss function is defined as:

$$L = -\sum_{i=1}^{T} y_i \log S_i \tag{8}$$

where y_i is the true value of the i^{th} category.

During the gradient descent of BiGRU, the gradient of the function storing the previous layer information is:

$$\frac{\partial L}{\partial a_c^t} = \frac{\partial L}{\partial h^t} a_z^t (1 - a_c^{t^2}) \tag{9}$$

where a_c^t is the unit storing the previous layer information, and h^t is the output value at time t. The gradient of the update gate is:

$$\frac{\partial L}{\partial a_z^t} = (-\frac{\partial L}{\partial h^t} h^{t-1} + \frac{\partial L}{\partial h^t} a_c^t)(1 - a_z^t) \tag{10}$$

The gradient of the reset gate is:

$$\frac{\partial L}{\partial a_r^t} = \frac{\partial L}{\partial a_c^t} h^{t-1} a_r^t (1 - a_r^t) \tag{11}$$

Therefore, the error gradient of the backpropagation of the network can be expressed as:

$$\frac{\partial L}{\partial h^{t-1}} = w_{ir}^t \frac{\partial L}{\partial a_r^t} + w_{iz}^t \frac{\partial L}{\partial a_z^t} + w_{ic}^t \frac{\partial L}{\partial a_c^t} a_r^t + \frac{\partial L}{\partial h^t}(1 - a_z^t) \tag{12}$$

IV. EXERIMENTAL RESULTS

To verify the effectiveness of the proposed model, the proposed algorithm is tested on a self-collected dataset. The experiments consist of six parts: the effectiveness verification of the model, the effect of the heart sounds on results, the influence of different time-frequency analysis methods on results, the effect of transfer learning on results, the influence of the retraining layer on results, and comparison of different methods. To test the generalization ability of the model, we use six-fold cross-validation in the experiment. The lung sound dataset is randomly divided into six subsets where the proportion of samples in each subset is the same for each category. 5 subsets are taken as the training data in turn, and one subset as the testing data. Finally, the mean of six times is taken as the final results.

For convenience, we abbreviate the lung sound recognition method based on VGGish-BiGRU model proposed in this paper as LSR-VBG. In experiments, the data preprocessing is completed on Matlab R2016a, and the construction and recognition of the deep transfer model is carried out on TensorFlow 1.8. GPUs used in experiments are NVidia P4 (four GPUs in total). The system is Linux Ubuntu 18.04 and the language is Python3.6. In the proposed model, the loss function of the model and the hyper parameters such as learning rate, iterative number, and batch size are jointly optimized by a large number of experiments, and finally the optimal network parameters are determined in the paper.

We obtained the optimal parameters by gradually tuning. The parameters are listed in table 1.

TABLE 1. List of parameters.

Parameter name	Parameter value
Learning rate	0.0001
Batch size	1
Optimizer	Adam
Iterative number	150
Regularization method	L2 regularization

A. DATASET

The experimental data were collected by the professional doctors using 3M Littmann 3200 electronic stethoscope. During collecting the lung sounds, each person was collected 30 seconds respiratory sound signals under the natural status. The normal lung sounds were collected from volunteers with the healthy lung status, and the abnormal lung sounds were collected from patients with pneumonia and asthma diseases in the hospital. The patient selected as samples include male and female. They are adults with different ages. And the severity of their lung disease is various. Some lung sound data are of poor quality because of uncorrected collecting method. It is difficult to train and test the model by these data. Therefore, we selected 384 typical lung sound samples with better quality from all collected lung sounds. They include 120 normal lung sounds, 156 pneumonia sounds and 108 asthma sounds. The frequency band of the lung sound signal is about 100HZ to 2000Hz. According to Nyquist sampling theorem, the sampling frequency of the experimental signal is set to 4000Hz.

In experiments, each lung sound is divided into three 10s segments to input the lung sound into the network. Thus, the dataset is expanded to include 360 normal lung sound samples, 468 pneumonia sound samples, and 324 asthma sound samples. The length of each sample is 10s. The total number of samples is 1152.

B. EFFECTVIENESS VERIFICATION OF THE MODEL

In experiments, the parameters of VGGish network provided by Google are directly transferred to the model and frozen. Then BiGRU network is fine-tuned by using the lung sound data. Finally the model is tested with the testing data. The results are shown in table 2.

TABLE 2. Results of LSR-VBG (%).

	Asthma	Pneumonia	Normal	Total
Precision	83.33	86.75	91.94	
Recall	85.17	89.23	87.11	87.41
F1-score	84.24	87.97	89.46	

From the results, the precision and F1-score of the normal lung sound is the highest, and the recall is lower than other two categories. The precision of pneumonia is lower than that of the normal lung sound, and its recall and F1-score are higher. The precision, recall and F1-score of asthma are all the lowest in three categories. The total accuracy reaches 87.41%, which demonstrates that the proposed model can better recognize the lung sounds. To further analyze the misclassification of various categories, the confusion matrix is used to analyze the experimental results in detail. The confusion matrix of the above results is given in table 3.

TABLE 3. Confusion matrix of results.

Real	Prediction			Total
	Asthma	Pneumonia	Normal	
Asthma	270	31	23	324
Pneumonia	36	406	26	468
Normal	11	18	331	360
Total	317	455	380	1152

From the confusion matrix, the probability of asthma recognized as pneumonia is higher than that of being recognized as the normal, and the probability of pneumonia recognized as asthma is also higher than that of being recognized as the normal. It indicates that pneumonia and asthma have a greater similarity from the view of the audio.

C. EFFECT OF THE HEART SOUNDS ON RESULTS

To test the influence of the heart sounds on the recognition results, the proposed model is firstly performed on the lung sound dataset which are not deleted the heart sounds. Then the model is executed on the lung sound dataset that has been removed the heart sounds. The results are shown in table 4.

TABLE 4. Effect of the heart sounds on results (%).

	Asthma	Pneumonia	Normal	Total
The heart sounds being not removed	65.74	76.50	86.11	76.78
The heart sounds being removed	83.33	86.75	91.94	87.41

It can be concluded from the experimental results that the recognition accuracy is greatly affected by the heart sounds. When the heart sounds are not removed, the recognition accuracy of three kind of the lung sounds is all lower than that from deleting the heart sounds from the lung sounds. The recognition accuracy of asthma reduces nearly 20%, the total accuracy drops about 10%. Experiments show that the heart sound is an important interference factor in the lung sound recognition. The recognition accuracy is greatly improved by using the wavelet threshold method to remove the heart sounds, especially the accuracy of asthma is greatly improved.

D. INFLUENCE OF DIFFERENT TIME-FREQUENCY ANALYSIS MEHTODS ON RESULTS

To compare the effects of different time-frequency analysis methods on the recognition results, STFT, MFCC, and Mel spectrogram are used to extract features from the lung sounds. The features are input into the deep transfer learning models proposed in the paper. The results are shown in table 5. Here, the window functions and the frame shift in the three time-frequency analysis methods are the same. Per second lung sound is divided into 96 frames, and each frame is extracted 64 features. The input of the model is a $10 \times 96 \times 64$ feature vector.

TABLE 5. Comparison of different time-frequency analysis methods (%).

Method	Asthma	Pneumonia	Normal	Total
STFT	78.09	85.47	88.06	84.20
MFCC	80.25	88.68	91.67	87.24
Mel spectrogram	83.33	86.75	91.94	87.41

From the results, the recognition accuracy obtained by STFT is lower than those of MFCC and Mel spectrogram. The accuracy from Mel spectrogram is near to that of MFCC. However, the difference of the accuracy of three classes from MFCC is bigger that from Mel spectrogram, especially the accuracy of asthma from MFCC is relatively poor.

Therefore, we select Mel spectrogram for time-frequency analysis in the following experiments.

E. IMPACT OF TRANSFER LEARNING ON RESULTS

To validate the effect of transfer learning on results, we compare the model including transfer learning with the model without transfer learning. In the two cases, the data collecting, the preprocessing, the feature extraction method, and the input dimension are the same. For without transfer learning, the model still uses the architecture adopted VGGish-BiGRU. The parameters of VGGish are randomly initialized, but are not transferred from the results obtained by Audio Set. The results are shown in table 6.

TABLE 6. Effect of transfer learning on results (%).

Method	Asthma	Pneumonia	Normal	Total
CRNNs	78.40	84.62	90.27	84.64
OURS	83.33	86.75	91.94	87.41

From the results, transfer learning effectively improves the experimental results. The recognition accuracy is improved, especially the accuracy of asthma. It shows that the lung sound recognition can be effectively improved by transferring the parameters from the source domain model.

At the same time, to verify the reliability of Audio Set as the source domain of transfer learning for the lung sound recognition, we compare Audio Set with two other open datasets. They are GTZAN dataset and GTZAN dataset. GTZAN dataset is a music audio dataset, which includes 10 kind of audio, such as jazz, rock, folk, pop, etc. Each category has 100 audio samples, and each audio is about 175 seconds. UrbanSound8K dataset is ambient sound dataset, which contains 10 kind of audio, such as car horn, children's play, dog bark, drill hole, etc. It contains 8732 audio samples, and each category includes a different number of audio. Each sample is about 4 seconds. In experiments, Audio Set is replaced with the two datasets as the source domain data. The preprocessing and the feature extraction method are the same as the previous experiments. When the two datasets are input into the network, they are resampled to the same frequency of Audio Set, namely 16000Hz. The frequency range is set to 125Hz-7500Hz while extracting Mel spectrogram. The environment and other processing are the same as the previous experiments. The results are given in table 7.

TABLE 7. Comparison of different source domains (%).

Source domain	Asthma	Pneumonia	Normal	Total
GTZAN	58.64	67.09	70.56	67.24
UrbanSound8K	62.96	72.86	73.89	70.40
Audio Set	83.33	86.75	91.94	87.41

From the results, the accuracy of each category is greatly decreased when GTZAN and UrbanSound8K are taken as the source domain dataset. And the difference between different categories is also relatively bigger. It demonstrates that it generates a certain negative transfer on the target data when GTZAN and UrbanSound8K are taken as the domain source dataset. There are two main reasons to lead to the results. One is that the data volume of the two datasets is much smaller than that of Audio Set. Another is that GTZAN and UrbanSound8K are both single domain sound data, which contains the number of audio categories much lower than that of Audio Set. The two sets are too unique and not universal. In this case, it will cause the phenomenon of negative transfer when transferring the knowledge to the target domain. According to the results, Audio Set has good generality and can be used as the source domain of transfer model for the lung sound data.

F. EFFECT OF RETRAINING METHODS ON RESULTS

During transferring the model parameters, although the transfer learning can bring the common features learned from the source domain data, there is only a few cardiopulmonary sound data in the source domain data. And there is less similarity between the target data and the source domain data. Therefore, it is necessary to retrain the model for extracting some unique features from the target dataset. It will better improve the recognition accuracy. To verify the importance of the retraining process, VGGish transfer network is directly connected to the flatten layer and the softmax layer for the classification after deleting the retraining layer. The results are given in table 8.

TABLE 8. Effect of retraining on results (%).

	Asthma	Pneumonia	Normal	Total
Without retraining	75.31	78.21	70.83	75.09
With retraining	83.33	86.75	91.94	87.41

From the results, the recognition accuracy is greatly reduced when deleting the retraining layer. Experiments show that the retraining can better improve the recognition accuracy of the target data when the source domain data and the target domain data are not much similar.

To test the performance of the retraining network BiGRU, it is compared with the fully connected layer, LSTM, GRU, and BiLSTM. In experiments, the number of nodes in the fully connected layer is 128. The hidden layer in LSTM and GRU includes 128 nodes. The number of nodes in the hidden layer in BiLSTM is 2 × 128. The input dimension of the fully connected layer is 10 × 128. Other parameters are the same as those in BiGRU. The results are shown in table 9.

TABLE 9. Comparison of different retraining methods (%).

Methods	Asthma	Pneumonia	Normal	Total
FC	78.40	78.85	74.44	77.34
LSTM	80.25	83.33	92.50	85.33
GRU	78.70	81.41	92.22	83.94
BiLSTM	79.63	84.62	94.44	86.28
BiGRU	83.33	86.75	91.94	87.41

From the above table, the recognition accuracy of the model is slightly improved by taking single fully connected layer as the retraining layer. The accuracy from taking single fully connected layer as the retraining layer is much lower than that from taking structural variants of the recurrent neural network as the retraining. The recognition accuracy of the normal lung sound from LSTM, GRU, and BiLSTM is slightly higher than that from BiGRU. The accuracy of the pneumonia, and asthma from BiGRU is higher than that from LSTM, GRU, and BiLSTM, especially BiGRU greatly improves the recognition accuracy of asthma. It shows that BiGRU as the retraining network has better balance than other RNN networks. And the mean accuracy from BiGRU is the highest in all cases. Therefore, we take BiGRU as the retraining part of the model.

G. COMPARISON WITH OTHER METHODS

To validate the effectiveness of the proposed algorithm, the model is compared with SVM, KNN, random forests (RF), 2L-CNN, 5L-CNN, and DNN-HMM. The results are shown in table 10. In experiments, Mel spectrogram features of the lung sounds are as the input of these methods. And Mel spectrogram is taken as a whole to input into the model instead of segmenting into 10 slices in SVM KNN, random forests, 2L-CNN, 5L-CNN, and DNN-HMM. The input dimension is 960 × 64. SVM, KNN, and Random Forests are executed by using Scikit-learn (Sklearn) library in Python. In SVM, the penalty parameter C is 0.9, the kernel function is Gaussian kernel function, and the rest parameters are the default value. In KNN, the number of the nearest neighbors is 6, the rest parameters use the default values. In random forests, the number of decision trees is 100. 2L-CNN [24] consists of two convolutional layers, two pooling layers, and a fully connected layer. The size of the convolution kernel is 3 × 3, and the two convolutional layers both contain 64 convolution kernels. The pooling layers adopt 2 × 2 max pooling. The fully connected layer includes 128 hidden nodes. Finally softmax is used to output the category of the lung sounds. The number of iterations is 350, the learning rate is 0.001, and the optimizer is Adam. 5L-CNN [25] includes five convolution layers and five max pooling layers. The sofmax layer is taken as the classification layer. The size of the first two convolution layers is 7*7 and 5*5, and that of the last three layers are 3*3. The learning rate is 0.01, and the dropout rate is 0.5. DNN-HMM [27] contains a dual hidden layer deep neural network and a hidden Markov model, where the number of the hidden nodes is 500 and the number of iterations is 150. The parameters in other six algorithms are the most optimal by gradually tuning.

TABLE 10. Comparison with other methods (%).

Methods	Asthma	Pneumonia	Normal	Total
SVM	77.16	81.84	88.61	82.64
KNN	66.67	72.22	76.67	72.05
RF	59.26	79.27	80.28	73.96
2L-CNN	61.11	73.08	79.44	71.70
5L-CNN	62.35	75.21	83.61	74.22
DNN-HMM	75.93	84.62	86.39	82.73
OURS	83.33	86.75	91.94	87.41

From the results, our method achieves the best recognition results. The main reason is that the model uses transfer learning to better solve the problem of insufficient data, and the model fully utilize the time series feature of the lung sounds by using BiGRU to retraining the network. In contrast to other algorithms, the proposed method greatly improves the recognition accuracy of three kind of the lung sounds, especially greatly progresses the accuracy of asthma.

V. CONCLUSION

The prevention and diagnosis of the lung diseases plays an important role in the human life system. Lung auscultation is one of the important methods for the lung disease detection. Most of recognition of lung sounds have weak generalization ability because they have higher dependency on data and the artificial features. Simultaneously, the recognition effect will be affected because the traditional CNN doesn't extract the temporal features of the lung sounds. To this end, a lung sound recognition model based on VGGish-BiGRU is proposed, which uses transfer learning and combines VGGish network with BiGRU network. BiGRU can capture the time series features of the lung sounds. Experiments show that the proposed algorithm effectively improves the recognition accuracy of the lung sounds, especially the accuracy of asthma in contrast to other methods. The method has higher generalization ability as well as better captures the temporal features of the lung sounds. At the same time, Experiments show that the low-frequency noise in the lung sounds is better deleted by high-pass filtering and the heart sounds in the lung sounds are better removed by the wavelet threshold method. And, the algorithm can be improved from the following aspects.

Firstly, lung sound samples are relatively few, which affects the accuracy of the model. And only one type of electronic stethoscope is used during collecting lung sound, which causes lack of data diversity. Moreover, the dataset only includes the normal lung sounds and two category lung disease samples. The model is needed to generalize other lung diseases.

Secondly, the model uses transfer learning. Although AudioSet has better generality, it has less similarity with the lung sounds, which also affect the recognized accuracy. The model needs to find more appropriate source domain to optimize the model.

In the future, we will collect more lung sounds, and further optimize the model. And the method is applied in auxiliary diagnosis of pulmonary diseases.

REFERENCES

[1] Z. Li and M. Du, "HHT based lung sound crackle detection and classification," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst.*, Dec. 2006, pp. 385–388.

[2] L. P. Malmberg, K. Kallio, S. Haltsonen, T. Katila, and A. R. A. Sovijärvi, "Classification of lung sounds in patients with asthma, emphysema, fibrosing alveolitis and healthy lungs by using self-organizing maps," *Clin. Physiol.*, vol. 16, no. 2, pp. 115–129, 1996.

[3] P. Bokov, B. Mahut, P. Flaud, and C. Delclaux, "Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population," *Comput. Biol. Med.*, vol. 70, pp. 40–50, Mar. 2016.

[4] A. Kandaswamy, C. S. Kumar, R. P. Ramanathan, S. Jayaraman, and N. Malmurugan, "Neural classification of lung sounds using wavelet coefficients," *Comput. Biol. Med.*, vol. 34, no. 6, pp. 523–537, 2004.

[5] S. Abbasi, R. Derakhshanfar, A. Abbasi, and Y. Sarbaz, "Classification of normal and abnormal lung sounds using neural network and support vector machines," in *Proc. 21st Iranian Conf. Elect. Eng. (ICEE)*, May 2013, pp. 1–4.

[6] G. D. Liu and J. Xu, "Neural network recognition algorithm of breath sounds based on SVM," *J. Commun.*, vol. 35, no. 10, pp. 218–222, 2014.

[7] F. Z. Göğü , B. Karlik, and G. Harman, "Classification of asthmatic breath sounds by using wavelet transforms and neural networks," *Int. J. Signal Process. Syst.*, vol. 3, no. 2, pp. 106–111, Dec. 2015.

[8] Y. Shi, Y. Li, M. Cai, and X. D. Zhang, "A lung sound category recognition method based on wavelet decomposition and BP neural network," *Int. J. Biol. Sci.*, vol. 15, no. 1, pp. 195–207, 2019.

[9] Y. Liu, C. M. Zhang, Y. H. Zhao, and L. Dong, "The feature extraction and classification of lung sounds based on wavelet packet multiscale analysis," *Chin. J. Comput.*, vol. 29, no. 5, pp. 769–777, 2006.

[10] S. Li and L. Yi, "Feature extraction of lung sounds based on bispectrum analysis," in *Proc. 3rd Int. Symp. Inf. Process.*, Oct. 2010, pp. 393–397.

[11] F. Jin, F. Sattar, and D. Y. T. Goh, "New approaches for spectro-temporal feature extraction with applications to respiratory sound classification," *Neurocomputing*, vol. 123, pp. 362–371, Jan. 2014.

[12] D. Emmanouilidou, E. D. Mccollum, D. E. Park, and M. Elhilali, "Computerized lung sound screening for pediatric auscultation in noisy field environments," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 7, pp. 1564–1574, Jul. 2018.

[13] M. M. Azmy, "Classification of lung sounds based on linear prediction cepstral coefficients and support vector machine," in *Proc. IEEE Jordan Conf. Appl. Elect. Eng. Comput. Technol. (AEECT)*, Nov. 2015, pp. 1–5.

[14] R. Palaniappan, K. Sundaraj, and S. Sundaraj, "A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals," *Bioinformatics*, vol. 15, no. 1, p. 223, Dec. 2014.

[15] N. Sengupta, M. Sahidullah, and G. Saha, "Lung sound classification using cepstral-based statistical features," *Comput. Biol. Med.*, vol. 75, pp. 118–129, Aug. 2016.

[16] K. Kosasih, U. R. Abeyratne, V. Swarnkar, and R. Triasih, "Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1185–1194, Apr. 2015.

[17] A. Haider, M. D. Ashraf, M. U. Azhar, S. O. Maruf, M. Naqvi, S. G. Khawaja, and M. U. Akram, "Separation and classification of crackles and bronchial breath sounds from normal breath sounds using Gaussian mixture model," in *Proc. Int. Conf. Neural Inf. Process.*, Cham, Switzerland: Springer, 2014, pp. 495–502.

[18] A. Rizal, R. Hidayat, and H. A. Nugroho, "Entropy measurement as features extraction in automatic lung sound classification," in *Proc. Int. Conf. Control, Electron., Renew. Energy Commun.*, Sep. 2017, pp. 93–97.

[19] X. J. Yao, H. Wang, and S. Liu, "Research on recognition algorithms of lung sounds based on genetic BP neural network," *Space Med. Med. Eng.*, vol. 29, no. 1, pp. 45–51, 2016.

[20] L. M. Santiago-Fuentes, S. Charleston-Villalobos, R. González-Camarena, M. Mejía-Ávila, H. Mateos-Toledo, I. Buendía-Roldan, and T. Aljama-Corrales, "A multichannel acoustic approach to define a pulmonary pathology as combined pulmonary fibrosis and emphysema syndrome," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Jul. 2017, pp. 2757–2760.

[21] M. A. Islam, I. Bandyopadhyaya, P. Bhattacharyya, and G. Saha, "Multichannel lung sound analysis for asthma detection," *Comput. Methods Programs Biomed.*, vol. 159, pp. 111–123, Jun. 2018.

[22] S. Charleston-Villalobos, G. Martinez-Hernandez, R. Gonzalez-Camarena, G. Chi-Lem, J. G. Carrillo, and T. Aljama-Corrales, "Assessment of multichannel lung sounds parameterization for two-class classification in interstitial lung disease patients," *Comput. Biol. Med.*, vol. 41, no. 7, pp. 473–482, Jul. 2011.

[23] A. Poreva, Y. Karplyuk, and V. Vaityshyn, "Machine learning techniques application for lung diseases diagnosis," in *Proc. 5th IEEE Workshop Adv. Inf., Electron. Electr. Eng.*, Nov. 2017, pp. 1–5.

[24] M. Aykanat, Ö. Kiliç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, pp. 65–73, Dec. 2017. .

[25] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," *Artif. Intell. Med.*, vol. 88, pp. 58–69, Jun. 2018.

[26] Q. Chen, W. Zhang, X. Tian, X. Zhang, S. Chen, and W. Lei, "Automatic heart and lung sounds classification using convolutional neural networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2017, pp. 1–4.

[27] L. Li, W. Xu, Q. Hong, F. Tong, and J. Wu, "Classification between normal and adventitious lung sounds using deep neural network," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Oct. 2017, pp. 1–5.

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

[29] J. F. Gemmeke, D. P. W. 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," *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 776–780.

[30] I. W. Selesnick and C. S. Burrus, "Generalized digital butterworth filter design," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1688–1694, Jun. 1998.

[31] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: https://arxiv.org/abs/1412.3555

LUKUI SHI received the B.S. degree in computer and application and the M.S. degree in computer application technology from the Hebei University of Technology, Tianjin, China, in 1996 and 1999, respectively, and the Ph.D. degree in computer application technology from Tianjin University, Tianjin, in 2006.

Since 2014, he has been a Professor with the School of Artificial Intelligence, Hebei University of Technology. He is the author of two books and more than 20 articles. His research interests include machine learning, lung sound recognition, and data digging.

Prof. Shi was a member of Discrete Intelligent Computing Professional Committee of Chinese Association of Artificial Intelligence and a member of Visual Big Data Professional Committee of China Society of Image and Graphics.

KANG DU received the B.S. degree in software engineering from the North University of China, Taiyuan, China, in 2016, and the M.S. degree in computer science and technology from the Hebei University of Technology, Tianjin, China, in 2019.

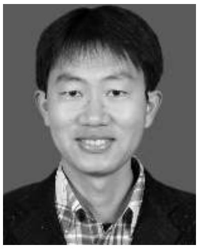
Since 2019, he has been an Engineer with SoundAI Technology Corporation, Limited. His research interests include lung sound recognition and audio recognition.

CHAOZONG ZHANG received the B.S. degree in computer science and technology from Lanzhou University, Lanzhou, China, in 2013, and the M.S. degree in computer technology from Shijiazhuang Tiedao University, Shijiazhuang, China, in 2016.

Since 2016, he has been an Engineer with the Hebei Institute of Scientific and Technical Information. His research interests include software information systems and big data application.

HONGQI MA received the B.S. degree in computer science and technology from Tangshan University, Tangshan, China, in 2017. She is currently pursuing the M.S. degree in computer science and technology with the Hebei University of Technology, Tianjin, China.

Her research interests include lung sound recognition and medical image processing.



WENJIE YAN received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China. He is currently an Associate Professor with the School of Artificial Intelligence, Hebei University of Technology. His research interests include data mining and machine learning.