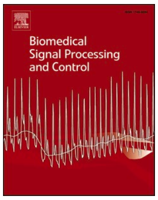
Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Research on lung sound classification model based on dual-channel CNN-LSTM algorithm

Yipeng Zhang [a,b], Qiong Huang [a], Wenhui Sun [a,b], Fenlan Chen [c], Dongmei Lin [d], Fuming Chen [a,*]

[a] *Medical Security Center, The 940th Hospital of Joint Logistics Support Force of Chinese People's Liberation Army, Lanzhou, Gansu 730050,China*
[b] *School of Information Engineering, Gansu University of Chinese Medicine, Lanzhou, Gansu 730000,China*
[c] *Lanzhou Rail Transit Co., Ltd, Lanzhou, Gansu 730051,China*
[d] *College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, Gansu 730050,China*

ARTICLE INFO

Keywords:
Lung sound classification
Mel cepstral coefficient
Convolutional neural network
Long short-term memory network

ABSTRACT

ulmonary diseases have a significant impact on human health and life safety, and abnormalities in the lungs are a direct response to lung diseases. Establishing an effective lung sound classification model that can assist in diagnosis is of great significance for electronic auscultation.In addressing the issue of lung sound signal classification, this study introduces a deep learning classification model based on a dual-channel CNN-LSTM algorithm. Initially, Mel-scale Frequency Cepstral Coefficients (MFCC) are employed for feature extraction from the dataset, transforming lung sound signals into Mel spectrograms. On this foundation, a dual-channel algorithm classification model is constructed, with parallel Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) modules. The CNN module is designed to capture spatial dimension features of the input data, while the LSTM module focuses on temporal dimension features. These two feature sets are fused together, enabling the model to classify lung sounds and thereby assisting in diagnosing pulmonary diseases for healthcare practitioners. This experiment used the ICBHI2017 Challenge Lungs dataset and obtained 5054 pieces of data through data augmentation and sampling techniques.The results show that the accuracy, recall, and F1 score of this model reach 99.01%, 99.13%, and 0.9915, respectively, significantly superior to other models, highlighting its practical application value.

1. Introduction

Respiratory diseases are non-communicable illnesses that can significantly impact the health and life safety of urban residents. In 2019, pulmonary diseases, represented by chronic obstructive pulmonary disease (COPD), pneumoconiosis, pulmonary nodules, and others, ranked fourth among the causes of death in our country's residents, accounting for 10.6 % of the national mortality rate [1]. Beyond posing a threat of mortality, respiratory diseases result in the loss and impairment of the patient's physiological functions, leading to additional occurrences of disability and increased treatment costs, placing a burden on both the individual's immediate family and society at large [2]. Apart from smoking, industrialization and the ongoing development of society are closely intertwined with the prevalence of respiratory diseases. The escalating severity of air pollution poses a growing risk, as inhaling toxic fumes or particulate matter can lead to severe pulmonary diseases [3].

Effectively diagnosing respiratory diseases in a timely manner is a crucial concern in the medical field. The pressing challenges posed by the recent COVID-19 pandemic have elevated lung sound detection technology to a current research hotspot [4]. Auscultation is one of the fundamental methods for diagnosing pulmonary issues, but traditional manual auscultation often introduces risks of misdiagnosis or omission for certain respiratory diseases. Additionally, it heavily relies on the subjective judgment of professional doctors, demanding a high level of expertise from healthcare practitioners.

Therefore, establishing an efficient lung sound classification model that aids in diagnosis is of paramount importance for clinical research. Such a model can analyze audio signals from the lungs to identify and differentiate various types of respiratory diseases, such as asthma, pneumonia, and chronic obstructive pulmonary disease. Traditional lung sound classification methods rely on manual feature extraction and simple machine learning algorithms, but these approaches suffer from low accuracy and unstable classification results. In recent years, the development of deep learning technology has significantly impacted the

[*] Corresponding author.
E-mail address: cfm5762@126.com (F. Chen).

https://doi.org/10.1016/j.bspc.2024.106257
Received 11 September 2023; Received in revised form 27 February 2024; Accepted 20 March 2024
Available online 25 March 2024

field of lung sound classification. Deep learning techniques can learn features directly from raw data, constructing highly expressive classification models to enhance accuracy and stability.

This paper proposes a Dual-channel Convolutional Neural Network and Long Short-Term Memory algorithm, incorporating the Tensorform layer, to improve lung sound classification accuracy. By establishing connections between lung sounds and respiratory diseases, the algorithm can identify the potential types of pulmonary diseases a patient may be suffering from, thereby enhancing the practical application of assistive electronic stethoscopes.

2. Relevant work

Traditional methods for classifying lung sounds are based on manual feature extraction and simple machine learning algorithms. Falk et al. [5] proposed representing cardiorespiratory sound signals using the time trajectory of short-time frequency spectrum energy. They applied bandpass and bandstop modulation filters for preprocessing cardiorespiratory sound signals. Ayari et al. [6] introduced an adaptive filtering algorithm by constructing multiple filter functions. MazicI et al. [7] developed a simple two-layer cascaded SVM pattern recognition architecture used for identifying wheezing sounds in respiratory recordings. ChenCH et al. [8] designed a digital stethoscope and a lung sound detection system that utilizes K-means clustering for feature clustering to identify different lung sounds. SenguptaN and team [9] proposed a novel feature set based on traditional cepstral feature statistical characteristics, employing an artificial neural network (ANN) for building an identification model. Zhang [10] suggested combining wavelet transform and BP neural network. However, these methods suffer from low accuracy and unstable classification results.

In recent years, the development of deep learning technology has advanced the field of lung sound classification. Deep learning techniques learn features directly from raw data, constructing highly expressive classification models to enhance accuracy and stability. Zhang [11] presented a lung sound classification model based on CNN-BiGRU. One-dimensional convolutional kernels in CNN preserve the temporal dimension of input features, allowing BiGRU to extract temporal features, effectively leveraging the advantages of spatial and temporal feature extraction. Choi Y et al. [12] proposed a lung disease classification model using attention modules and deep learning, extracting respiratory sounds using log-Mel spectrogram MFCC. Although effective in classifying normal and five types of adventitious sounds, the model overlooks the importance of spatial features by focusing only on the temporal characteristics of lung sound signals. Petmezas G [13] initially used features extracted from short-time Fourier transform spectrograms by a convolutional neural network (CNN) as input to a long short-term memory (LSTM) network for classifying four types of lung sounds. Despite achieving only 75.57 % accuracy through ten-fold cross-validation, the feature extraction and model design require improvement for enhanced accuracy.

Therefore, this paper proposes a dual-channel neural network model based on CNN and LSTM for classifying lung sound signals. Data processed through MFCC is simultaneously fed into the CNN and LSTM network layers, fully utilizing the spatial feature extraction capabilities of CNN and the temporal information feature extraction capabilities of LSTM. Extracted feature information is then fused and input into the classifier for final classification.

3. Dual-Channel CNN-LSTM model

3.1. Mel-Frequency cepstral coefficients

For lung sound signals, feature extraction techniques are developed to extract data that can differentiate between different types of lung sound signals from various signal data. When it comes to lung sound signals, it is crucial to use appropriate transformation methods to quantify the features of different lung sound signals in the time and frequency domains for classification. Therefore, feature extraction methods have a direct impact on the distinctiveness of features and classification accuracy for different lung sound signals. Altan [14], for instance, used statistical features based on the Hilbert-Huang Transform (HHT) of lung sounds and deep learning algorithms to distinguish between COPD patients and healthy subjects. However, the accuracy of this method still needs improvement. Commonly used audio feature parameters include Linear Predictive Cepstral Coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) [15]. In this paper, MFCC is proposed for feature extraction from lung sound signals. MFCC is a commonly used feature extraction method in speech recognition. It involves extracting cepstral parameters in the Mel-scale frequency domain. The Mel scale describes the nonlinear characteristics of human hearing [16]. It is based on the characteristics of the human auditory system in perceiving sounds in various frequency ranges [17]. The relationship between the Mel scale and frequency can be approximated as shown in Eq. (1):

$$\mathrm{Mel}(f) = 2595 \times \lg\left(1 + \frac{f}{700}\right) \tag{1}$$

The process of extracting MFCC involves preprocessing, Fast Fourier Transform (FFT), a set of Mel filters, logarithmic operations, and Discrete Cosine Transform (DCT), among other steps. The extraction process is illustrated in Fig. 1.

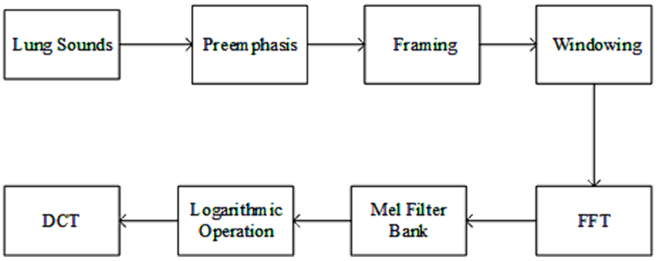


Fig. 1. Process of MFCC Feature Extraction.

The purpose of pre-emphasis is to preprocess the original signal by applying a high-pass filter, emphasizing the high-frequency components, and reducing the amplitude of the low-frequency components to enhance the effectiveness of subsequent analysis. The result after pre-emphasis processing can be represented by Eqs.(2):

$$y(n) = x(n) - ax(n-1) \tag{2}$$

In this Eqs, $x(n)$ represents the sample value at time n, and a is the pre-emphasis coefficient, which is set to 0.98 in this case.The pre-emphasized signal is then divided into short-time frames, typically with a duration of 20–40 ms. This is done because lung sound signals exhibit short-time stability over time, and segmenting the signal into frames assumes that the signal is relatively stable within each time segment.A window function, such as the Hanning or Hamming window, is applied to each frame.The purpose of the window function is to reduce discontinuities at the frame boundaries and prevent spectral leakage.By multiplying each frame with the Hamming window, a discrete Fourier transform (DFT) is applied to obtain the energy distribution in the frequency domain for each windowed frame [18]. Next, a Fast Fourier Transform (FFT) is applied to each windowed frame, converting the time-domain signal into a frequency-domain representation. The FFT transforms each frame into spectral information, providing the energy distribution of the frame in the frequency domain, as shown in Eq. (3):

$$\mathrm{x_a}(k) = \sum_{n=0}^{N-1} \mathrm{x}(n) \times exp\left(\frac{-j2\pi k}{N}\right), \quad 0 \leqslant k \leqslant N \tag{3}$$

By convolving the power spectral density of each frame with a set of Mel filters, it's possible to reduce data volume and computational complexity. The output of the filterbank can be represented as in Eq. (4):

$$\mathrm{H_m}(k) = \begin{cases} 0, & k\leqslant f(m-1) \\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m-1)\leqslant k\leqslant f(m) \\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m)\leqslant k\leqslant f(m+1) \\ 0, & k\geqslant f(m+1) \end{cases} \tag{4}$$

Taking the logarithm of the filterbank's output, as shown in Eqs. (5), allows for energy compression. This is because the human perception of sound intensity is nonlinear, and applying a logarithmic compression to energy values better approximates the way the human ear perceives sound.Specifically, the human ear is more sensitive to low-frequency sounds and less sensitive to high-frequency sounds [19]:

$$S(m) = \ln\left(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)\right), 0\leqslant m\leqslant M \tag{5}$$

The discrete cosine transform (DCT) is applied to the signal $S(m)$after logarithmic compression to convert the spectrum into cepstral coefficients $C(n)$, as described in Eq. (6):

$$C(n) = \sum_{m=0}^{N-1} S(m) \times \cos\left(\frac{\pi n(m-0.5)}{M}\right), \quad n = 1,2,3....L \tag{6}$$

The Lth order refers to the order of MFCC coefficients; M represents the number of triangular filters. Cepstral coefficients constitute the final feature representation of MFCC and possess good discriminative capabilities.

3.2. Convolutional neural network

The Convolutional Neural Network (CNN) is a deep learning model extensively utilized in image recognition. It mimics the functionality of the visual cortex in biology, enabling automatic feature extraction and classification of two-dimensional data such as images. A complete two-dimensional CNN typically consists of modules including input layers, convolutional layers, pooling layers, fully connected layers, and output layers [20]. The core concept of CNN involves constructing a network structure using convolutional layers, pooling layers, and fully connected layers to extract features from images and perform classification. The architectural diagram of a convolutional neural network is depicted in Fig. 2.

The Convolutional Layer is a core component of CNN, composed of multiple convolutional kernels. Each convolutional kernel slides over the input image, performing a convolution operation to extract local features from the image. The convolution operation is achieved by element-wise multiplication and accumulation of the convolutional kernel with the corresponding positions of the input image, resulting in an output feature map. By using different convolutional kernels, this layer can detect low-level features such as edges and textures in the image. Additionally, the weight parameters within the convolutional layer are automatically learned from training data. The calculation process of the convolutional kernel is represented by Eq. (7):

$$y_j^k = f(k\sum_{i\in c_j} x_i^{k-1} * u_{ij}^k + b_j^k) \tag{7}$$

In the Eqs, * represents the convolution operation, and $f(x)$ represents the activation function, typically using the ReLU function or the sigmoid function. The role of the activation function is to perform a nonlinear mapping on the output of the convolutional layer, enabling the network to learn more complex features.

Following the convolutional layer, there is a Pooling Layer, which is used to reduce the spatial dimensions of the output from the convolutional layer, reducing the number of parameters and improving computational efficiency. Pooling operations also provide some degree of translation invariance, making the network more robust to small shifts in the input. The commonly used pooling operation is Max Pooling, which selects the maximum value in each region as the output after pooling, preserving important features while reducing the size of the image.

Typically, after the last pooling layer, Fully Connected Layers are added. Fully connected layers connect the feature maps from the pooling layers to a classifier for the final classification operation. In a fully connected layer, each neuron is connected to every neuron in the previous layer, and each connection has a weight for learning feature combinations. Fully connected layers usually consist of one or more fully connected hidden layers and an output layer.

3.3. Long Short-Term Memory neural network

Long Short-Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN)[21,22]. LSTM is designed to address the issues of vanishing gradients and exploding gradients that traditional RNNs encounter when dealing with long sequences [23,24]. This design allows LSTM to capture long-term dependencies more effectively, addressing both short-term and long-term dependencies [25,26]. The main

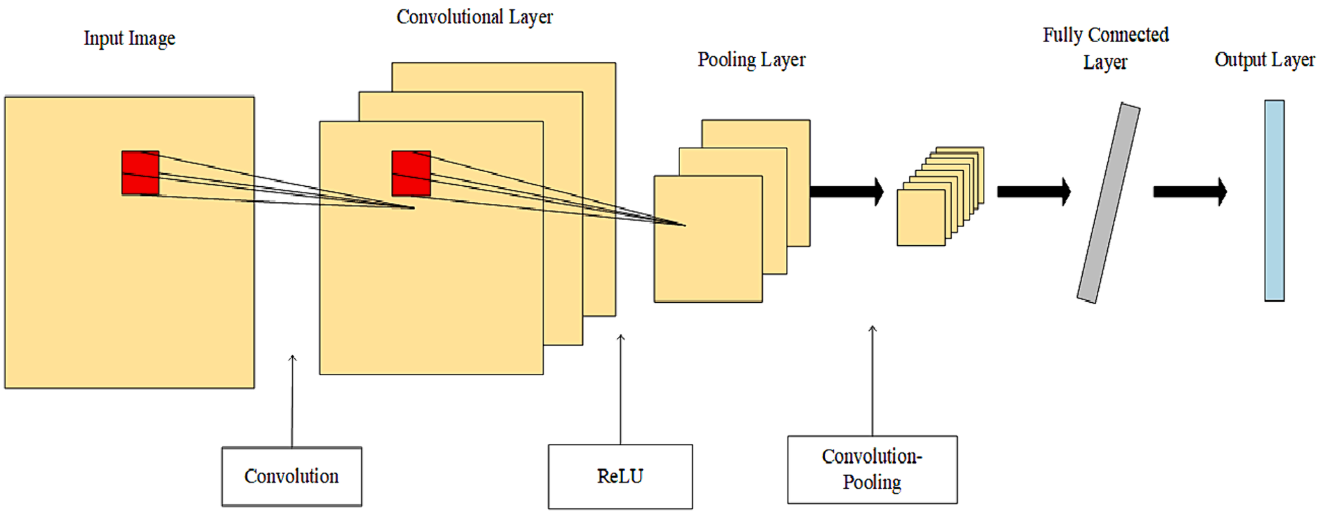


Fig. 2. Structural Diagram of Convolutional Neural Network Model.

distinction between LSTM and regular RNN lies in its internal structure. LSTM uses structures called "gates" to control the flow of information, thus effectively handling long-term dependencies. The network architecture of LSTM is illustrated in Fig. 3.

In Fig. 3, LSTM consists of three key components: the Input Gate, the Forget Gate, and the Output Gate. The Forget Gate determines which information should be discarded from the cell state. It reads the previous output ht-1 and the current input x_t, and computes an output f_t using a sigmoid activation function (σ). This output is then multiplied with the cell state C_{t-1} to control which information should be retained and what should be forgotten. This enables LSTM to remember relevant information from the past while discarding less important details. The computation is represented by Eq. (8):

$$f_t = \sigma(W_{fh}h_{t-1} + W_{fx}x_t + b_f) \tag{8}$$

The Input Gate controls the input of new information. It determines the update it using a sigmoid activation function and creates a new candidate value vector Q_t using a hyperbolic tangent (tanh) function. Finally, it uses element-wise multiplication and addition to decide how much new information should be added to the current cell state C_t. The calculation is represented by Eq. (9):

$$\begin{aligned} \mathrm{i}_t &= \sigma(W_{ih}h_{t-1} + W_{ix}x_t + b_i) \\ Q_t &= \tanh(W_{ch}h_{t-1} + W_{cx}x_t + b_c) \\ C_t &= f_t{}^*C_{t-1} + i_t{}^*C_t \end{aligned} \tag{9}$$

The Output Gate is used to determine which parts of the output hidden state will be passed on to the next time step. First, it controls the output weights O_t using a sigmoid activation function. Then, the cell state is processed through a tanh function (resulting in a value between −1 and 1), and this result is multiplied by the output of the sigmoid gate. The final output represents the part of C_t that is selected. The calculation is represented by Eq. (10):

$$\begin{aligned} o_t &= \sigma(W_{oh}h_{t-1} + W_{ox}x_t + b_o) \\ h_t &= o_t{}^*\tanh(C_t) \end{aligned} \tag{10}$$

3.4. Dual-Channel CNN-LSTM pulmonary sound classification model

In previous research, most scholars applying Convolutional Neural Networks (CNN) for lung sound classification focused solely on the temporal characteristics of lung sound data, neglecting the spatial properties between data points. This oversight may result in incomplete feature information extraction. Pan et al. [27], in their study on rolling bearing fault diagnosis, utilized the output of CNN as input for LSTM to identify bearing fault categories. However, this model disrupts temporal features because the resolution of the feature data extracted by CNN is lower than the original data. This disruption leads to the disturbance of the time-series characteristics of some data and the loss of certain temporal information from the original data. Consequently, when utilizing LSTM for further data feature extraction, it fails to fully utilize all the information from the original lung sound data and maintain complete temporal characteristics.

Therefore, this paper proposes a dual-channel feature fusion CNN-LSTM lung sound classification model that simultaneously considers both spatial and temporal characteristics of the data. The main components of this model include Long Short-Term Memory Neural Network (LSTM), Convolutional Neural Network (CNN), feature fusion layer, and output layer, as illustrated in Fig. 4.

One of the channels utilizes CNN to extract the "spatial" characteristics of the data, while the other channel employs LSTM to extract the "temporal" characteristics of the data. In contrast to the single-channel CNN-LSTM model, where LSTM operates on data that has been through CNN convolution and pooling and has lost the original data's temporal sequence features, here we use LSTM to directly process MFCC feature data to capture the temporal information features of the raw data. The role of the feature fusion layer is to merge the temporal and spatial characteristics of the pulmonary sound data, and the output layer determines the corresponding types based on the fused features.

To reduce training time and mitigate the issues of vanishing and exploding gradients, this model employs the backpropagation algorithm, computes the loss using the cross-entropy loss function, and uses the Adam optimizer for model optimization. The specific model parameter settings are shown in Table 1.

The main steps in establishing a lung sound classification model based on dual-channel CNN-LSTM are as follows:

(1) Data Split and Feature Extraction: The original lung sound data samples are randomly divided into training and testing sets. The MFCC features are extracted from the audio using the librosa library. MFCC converts lung sound signals into mel-frequency cepstral coefficient representations.
(2) Parameter Configuration: Using MFCC, the lung sound signals are transformed into mel spectrograms. These are then separately input to the LSTM and CNN layers for feature learning. Suitable

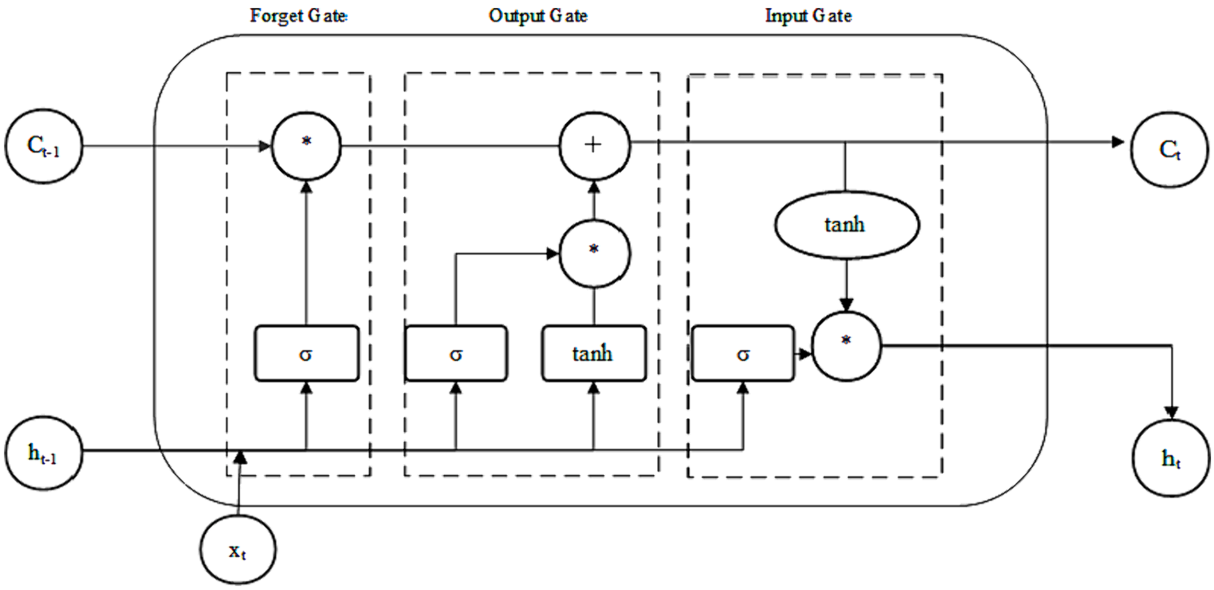


Fig. 3. Structure Diagram of Long Short-Term Memory Neural Network.

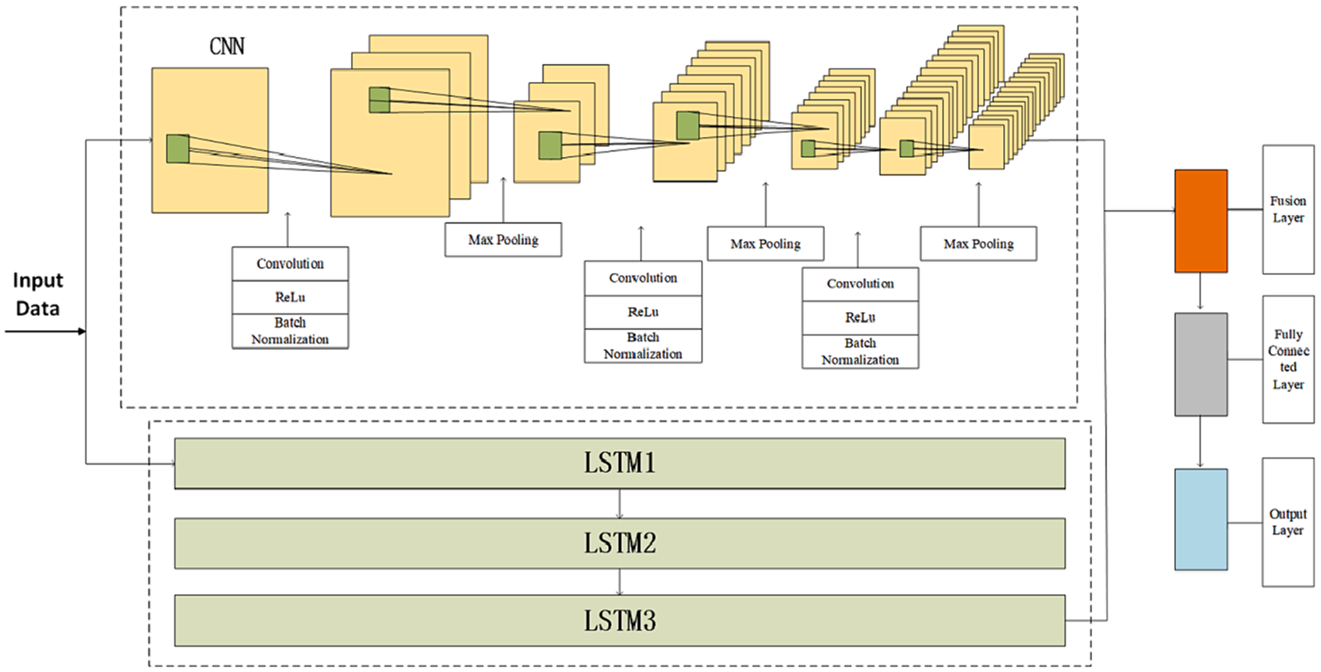


Fig. 4. Structure Diagram of the Dual-Channel CNN-LSTM Model.

Table 1
Structural Parameters of the Dual-Channel CNN-LSTM Model.

Layer	Number of Convolutional Kernels	Kernel Size	Number of Hidden Nodes	Output Dimension	Activation Function
Input Layer	–	–	–	(20,862,1)	–
Convolutional Layer 1	32	3*3	–	(20,862,32)	ReLU
Max Pooling Layer 1	–	2*2	–	(10,431,32)	
Convolutional Layer 2	64	3*3	–	(10,431,64)	ReLU
Max Pooling Layer 2	–	2*2	–	(5,215,64)	
Convolutional Layer 3	128	3*3	–	(2,107,128)	ReLU
Max Pooling Layer 3	–	2*2	–	(1,53,128)	
Flatten Layer	–	–	–	(1,6784)	–
LSTM1	–	–	512	(1,512)	Tanh, Sigmoid
LSTM2	–	–	256	(1,256)	Tanh, Sigmoid
LSTM3	–	–	128	(1,128)	Tanh, Sigmoid
Feature Concatenation	–	–	–	(1,6912)	–
Fully Connected Layer	–	–	–	(6912,6)	–

parameters such as learning rate, convolutional kernel size, LSTM unit count, and training iteration count are selected to ensure efficient learning of the model. Both CNN and LSTM are employed to extract spatial and temporal feature information from the lung sound data.

(3) Feature Fusion: The two distinct feature vectors obtained from step (2) are concatenated and fused. This results in a feature vector that contains non-periodic spatial features and periodic temporal features, effectively combining the two.

(4) Fully Connected Layer Classification and Parameter Optimization: The fused feature vector is input into a fully connected layer for the classification task. Simultaneously, model parameters are fine-tuned based on the trends in validation set loss values and accuracy, aiming to further enhance classification performance.

(5) Model Validation and Evaluation: K-fold cross-validation is used to select the optimal model. The testing set data is fed into the model, and based on the output results, classification metrics such as accuracy, precision, recall, and F1-score are calculated. The model's performance is evaluated accordingly.

4. Experimental validation and result analysis

The experimental data processing environment for this study included the Windows 11 operating system, PyTorch 1.8.1, Python 3.8, the CUDA 11.1 deep learning framework, a 14 vCPU Intel(R) Xeon(R) Gold 6330 CPU @ 2.00 GHz processor, and an RTX 3090 (24 GB) graphics card.

4.1. *Experimental data and processing*

4.1.1. *Experimental dataset*

The dataset used in this experiment is the ICBHI2017 Challenge Dataset [28]ICBHI2017 is a publicly available dataset for the analysis of respiratory sound signals. It includes respiratory sound signals from patients with various respiratory diseases as well as healthy individuals. This dataset was independently collected by two research groups from different countries over several years. The database consists of a total of 5.5 h of recordings from 920 subjects, comprising 126 annotated audio samples. These recordings were acquired using different devices and have varying durations ranging from 10 s to 90 s. Additionally, information about the chest location from which the recordings were obtained is provided. Respiratory experts in the field have annotated these lung sounds, and it includes a total of 8 types: Asthma, Bronchiectasis, Bronchiolitis, Chronic Obstructive Pulmonary Disease (COPD), Healthy, Lower Respiratory Tract Infection (LRTI), Pneumonia, and Upper Respiratory Tract Infection (URTI). The specific quantity analysis is depicted in Fig. 5.

4.1.2. Data preprocessing

As shown in Fig. 5, the dataset exhibits significant class imbalance, with varying quantities for each category. Particularly, two disease categories, Asthma and LRTI, have only 1 and 2 samples respectively. To address this imbalance, these two categories were removed, resulting in a final dataset of 6 classes with 917 samples. Given the relatively small dataset size and the pronounced class imbalance, data preprocessing techniques were applied, including data augmentation and sampling.

Data augmentation increases the diversity of training data and helps reduce overfitting. For the audio signals in this experiment, the data augmentation techniques used include:

(1) Time stretching: Adding silence or repeated segments to the original lung sound signal to increase its time duration.
(2) Pitch shifting: Altering the pitch or tone of the audio signal to introduce frequency variations.
(3) Background noise: Adding background noise to the lung sound signal to simulate environmental noise in real-world scenarios. This was achieved by generating white noise and applying a first-order low-pass Butterworth filter to add the noise signal to the lung sound.
(4) Data clipping: Randomly selecting a segment from the original lung sound signal to create lung sound segments of varying lengths.

To address the class imbalance issue, oversampling and undersampling techniques were employed to adjust the class distribution in the dataset. Oversampling involved performing additional random augmentations on the smaller classes, while undersampling involved random sampling from the augmented data of the larger classes. This helped achieve a more balanced distribution of samples. Fig. 6 illustrates the data distribution after preprocessing, showing a significant improvement in class balance compared to the original data distribution.

After data augmentation and sampling of the original data, a total of 5054 data instances were obtained. The dataset was randomly divided into three categories: training set, validation set, and test set. The split ratio for the combined training and validation sets to the test set was 8:2, and within the training set and validation set, the split ratio was also 8:2. The detailed partitioning is presented in Table 2.

4.2. K-Fold Cross-Validation for model selection

K-Fold Cross-Validation is a valuable technique in deep learning for selecting the best model. In deep learning, it’s essential to assess a model’s performance to determine whether it can generalize effectively to unseen data. The simplest approach is to split the dataset into training, validation, and test sets, where the training set is used for model training, the validation set is used to assess the model’s performance and adjust hyperparameters, and the test set is used for the final evaluation of the model’s performance. However, this approach has a randomness issue. If you use a single random data split to train the model, the results may be influenced by that specific split. K-Fold Cross-Validation helps better understand a model’s performance on different data subsets, reducing evaluation bias caused by random data splits and enhancing the reliability of model evaluation.

First, the training set in the dataset is divided into K folds, with each fold containing roughly the same number of samples. In this study, K is set to 5. This division is random, ensuring that each fold is independent and non-overlapping. One fold is chosen as the validation set, and the remaining four folds are combined into a training set. The classification model is trained on the training set and evaluated using the validation set. After completing 5 iterations, you will have 5 performance metrics, with each metric corresponding to a model performance evaluation on a validation set. The best model is selected, and the test set is used for final testing.

In this study, K-Fold Cross-Validation is used for model selection, with K set to 5. Table 3 shows the accuracy (Acc) and loss rate (Loss) obtained on each validation set for the four models in each fold.

For the results in Table 3, we choose the best models by considering both accuracy and loss rate. We select the 1st group of CNN, the 1st group of LSTM, the 5th group of Single-Channel CNN-LSTM, and the 5th group of Dual-Channel CNN-LSTM as the optimal models.

From Fig. 7, it can be observed that all models converge quickly and

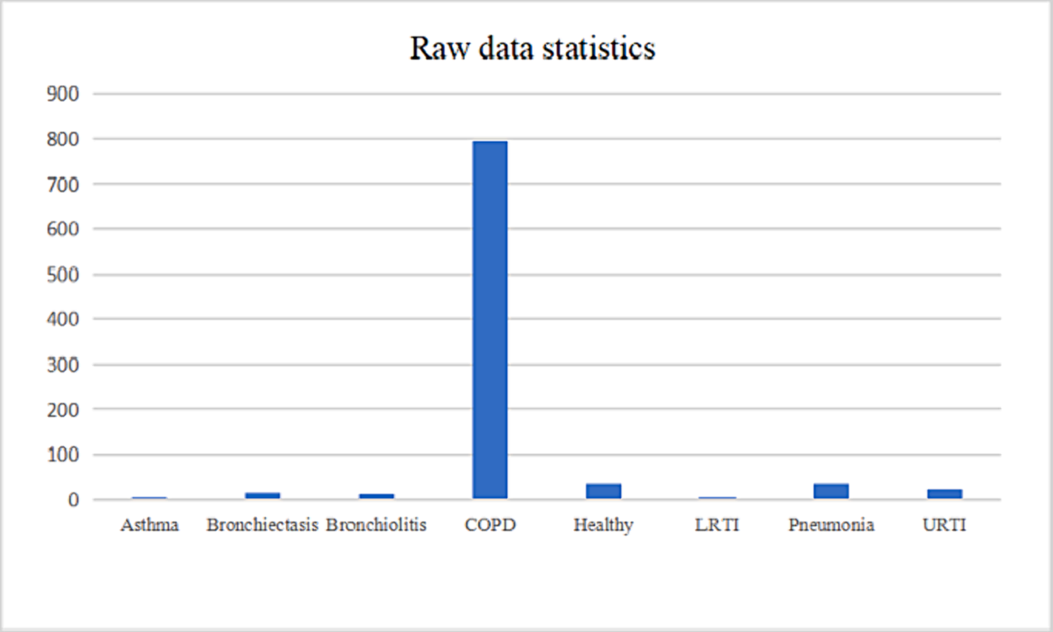


Fig. 5. Statistical Chart of Raw Data.

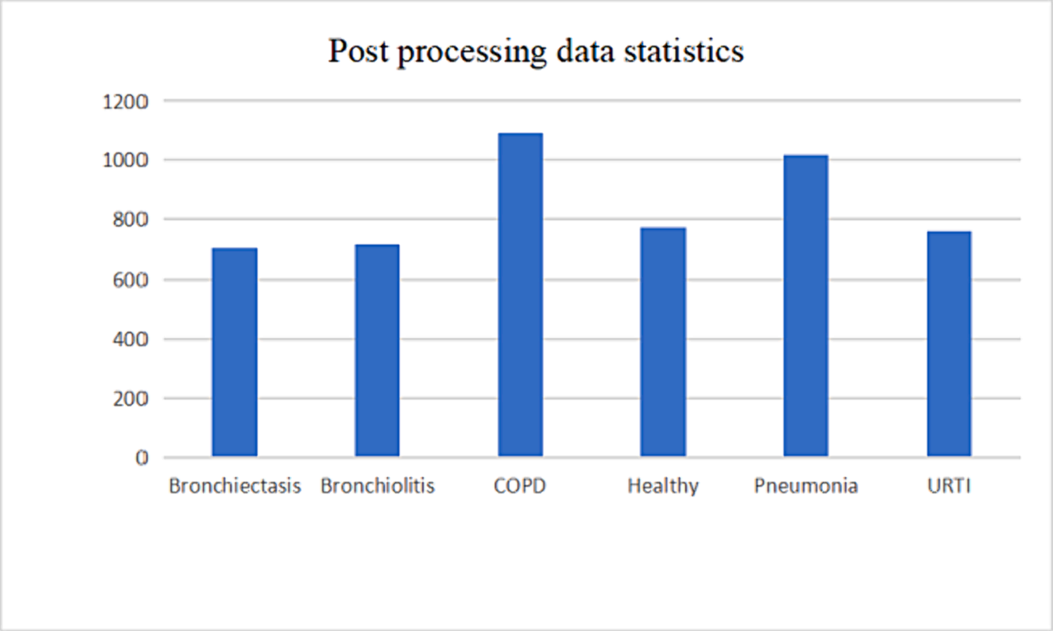


Fig. 6. Data Statistics after Preprocessing.

Table 2
Experimental Dataset Partitioning.

Disease name	Total quantity (piece)	Training set (pieces)	Verification set (piece)	Test set (pieces)
Bronchiectasis	704	454	114	136
URTI	715	463	116	136
Pneumonia	1090	701	175	214
Healthy	770	490	122	158
COPD	1017	646	162	209
Bronchiolitis	758	482	120	156

Table 3
Accuracy and Loss Rates on the Validation Sets for Each Model.

Model	Fold	Acc	Loss
CNN	1	97.12 %	13.88 %
	2	95.63 %	21.41 %
	3	94.84 %	22.77 %
	4	95.24 %	24.37 %
	5	96.53 %	17.76 %
LSTM	1	95.93 %	26.34 %
	2	94.64 %	39.13 %
	3	95.04 %	37.52 %
	4	95.63 %	28.43 %
	5	95.83 %	30.60 %
Single-Channel CNN-LSTM	1	98.11 %	12.32 %
	2	98.01 %	11.32 %
	3	97.82 %	14.81 %
	4	98.61 %	9.28 %
	5	98.71 %	8.62 %
Dual-Channel CNN-LSTM	1	98.71 %	5.92 %
	2	98.61 %	7.65 %
	3	98.51 %	6.21 %
	4	98.90 %	5.10 %
	5	99.03 %	3.88 %

stabilize. However, considering a combination of accuracy and loss rate, the dual-channel CNN-LSTM model exhibits the best performance on the dataset.

4.3. *Feature Visualization*

In order to test the performance of the model, under the same test set, the proposed model is compared with the single-channel CNN-LSTM model, CNN model, and LSTM model selected using K-fold cross-validation. When the initial parameters for learning rate and number of iterations are 0.0001 and 300 respectively, to compare the feature extraction performance of these models, the feature information extracted by CNN, LSTM, single-channel CNN-LSTM, and dual-channel CNN-LSTM is obtained. The t-distributed stochastic neighbor embedding (t-SNE) technique is used for visualizing the feature extraction, and the specific results are shown in Fig. 8.

The Fig. shows that the features extracted by the CNN model are distributed more randomly in space, with a lot of overlap between features. This indicates that the feature extraction capability of the CNN model is relatively weak and cannot clearly distinguish between different categories, leading to classification errors.While the LSTM model and the single-channel CNN-LSTM model have relatively distinct feature boundaries, their features are more scattered in space. This suggests that their feature extraction is relatively more dispersed, which may lead to classification errors in cases where some categories are similar.In comparison to the previous three models, the proposed dual-channel CNN-LSTM model in this paper exhibits more pronounced feature boundaries in the t-SNE plot, and features from the same category are more tightly clustered. This indicates that this model can better distinguish between different categories, and the extracted features are easier to differentiate, thus improving classification accuracy. Its features are more distinctive, making it more suitable for handling class imbalances or high similarity between classes in practical applications.

4.4. *Performance evaluation*

In order to comprehensively evaluate the classification performance of the network models, accuracy is first adopted as an evaluation metric. Accuracy is a convenient metric that allows for an intuitive comparison of the performance of different models. Next, we choose to use a confusion matrix to provide a detailed representation of the network model’s predictions for each class [29]. The confusion matrix comprehensively records the discrepancies between the model’s predicted results and the true values. The representation of the confusion matrix is

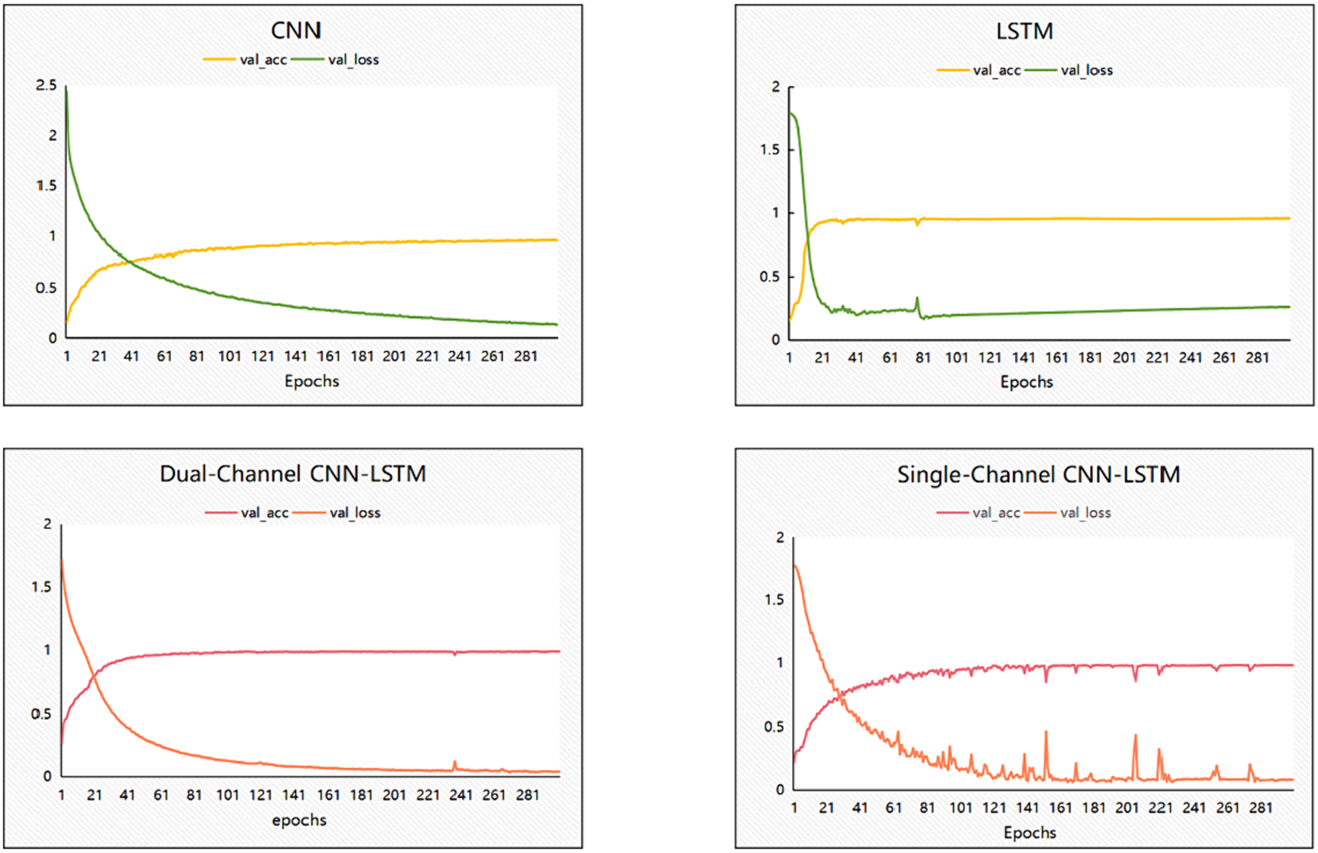
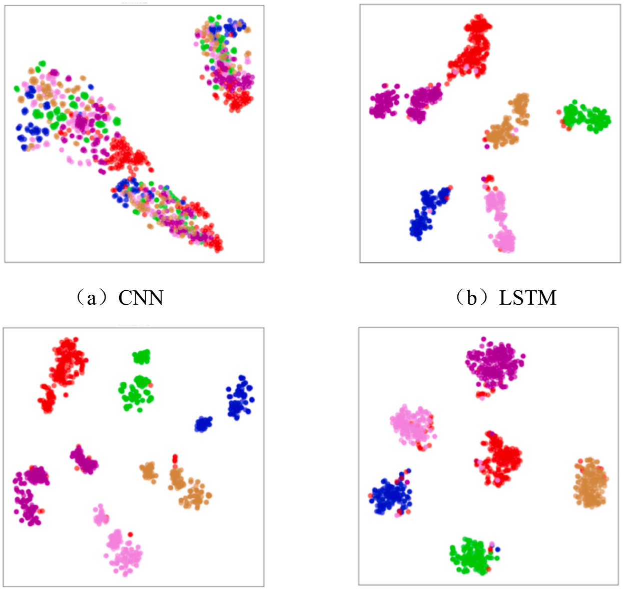


Fig. 7. Model accuracy and loss rate curve.



（a）CNN

（b）LSTM

（c）Single-Channel CNN-LSTM

（d）Dual-Channel CNN-LSTM

Fig. 8. Feature Visualization.

shown in Table 3. Taking the example of binary classification between normal and abnormal health conditions, where TP represents instances the model predicts as normal and they are indeed normal; FP represents instances the model predicts as normal, but they are actually abnormal; FN represents instances the model predicts as abnormal, but they are actually normal; TN represents instances the model predicts as abnormal and they are indeed abnormal.

Based on the confusion matrix results, secondary metrics can be derived, including precision, recall, and specificity. The calculation methods for these evaluation metrics are shown in Table 4.

In the actual process of lung sound classification, we face the presence of multiple diseases. Therefore, when using a neural network model for classification, it is essential to ensure that the classification results are neither missed nor duplicated, making recall a crucial evaluation metric. In addition, to avoid conflicts between evaluation metrics and make comprehensive assessments more manageable, this paper introduces the F1 score as an evaluation metric.

The F1 score takes into account both the recall and specificity of the neural network model. The F1 score has a range between 0 and 1, with higher scores indicating a stronger recognition ability of the network model across different samples. The calculation of the F1 score is shown in Eqs. (11).

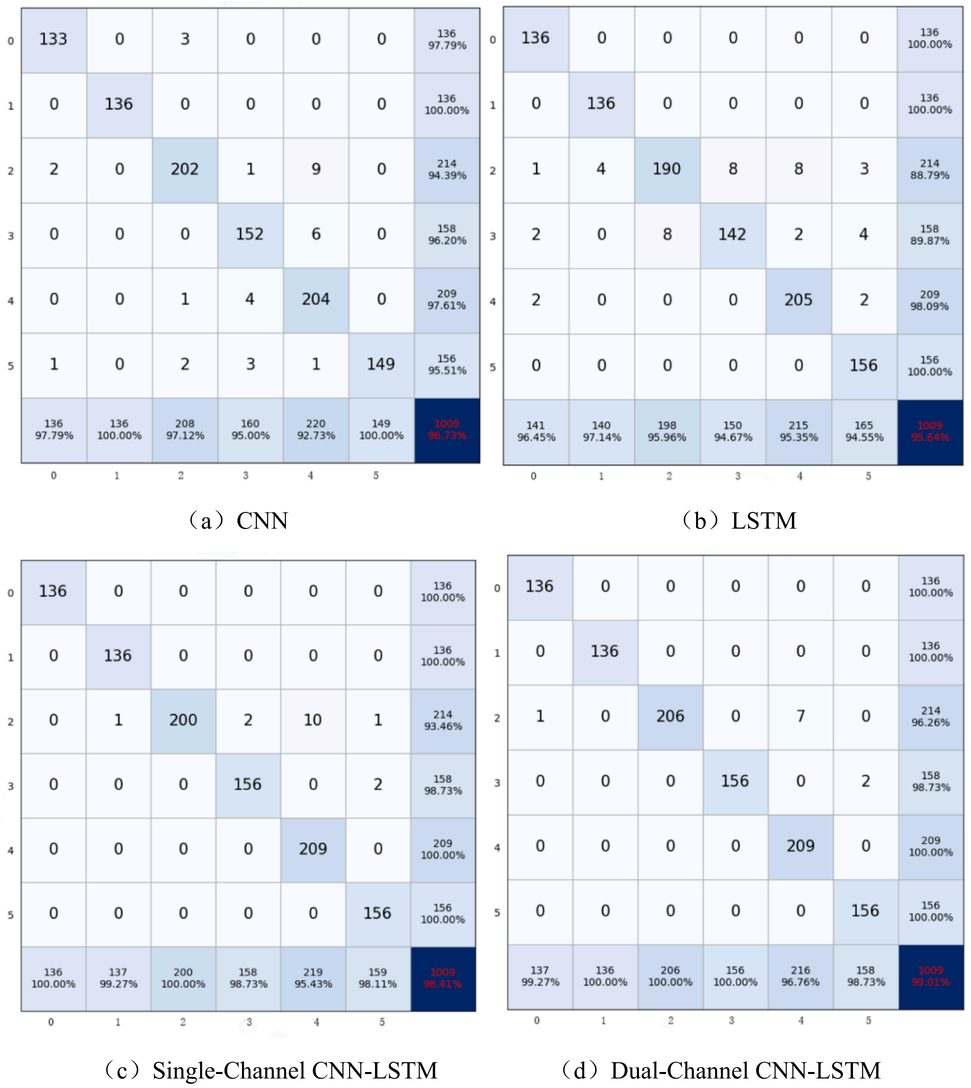
$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{11}$$

The confusion matrices for the four models, tested using the test dataset, are shown in Fig. 9. The classification results for the four models are presented. Among them, the proposed dual-channel CNN-LSTM model achieved the highest accuracy, reaching 99.01 %.

The confusion matrix results provide insights into the accuracy of

Table 4
Confusion Matrix.

Confusion Matrix		Real Type	
		Normal	Abnormal
Predicted Type	Normal	TP	FP
	Abnormal	FN	TN



	0	1	2	3	4	5	
0	133	0	3	0	0	0	136 97.79%
1	0	136	0	0	0	0	136 100.00%
2	2	0	202	1	9	0	214 94.39%
3	0	0	0	152	6	0	158 96.20%
4	0	0	1	4	204	0	209 97.61%
5	1	0	2	3	1	149	156 95.51%
	136 97.79%	136 100.00%	208 97.12%	160 95.00%	220 92.73%	149 100.00%	1009 96.73%

（a）CNN

	0	1	2	3	4	5	
0	136	0	0	0	0	0	136 100.00%
1	0	136	0	0	0	0	136 100.00%
2	1	4	190	8	8	3	214 88.79%
3	2	0	8	142	2	4	158 89.87%
4	2	0	0	0	205	2	209 98.09%
5	0	0	0	0	0	156	156 100.00%
	141 96.45%	140 97.14%	198 95.96%	150 94.67%	215 95.35%	165 94.55%	1009 95.64%

（b）LSTM

	0	1	2	3	4	5	
0	136	0	0	0	0	0	136 100.00%
1	0	136	0	0	0	0	136 100.00%
2	0	1	200	2	10	1	214 93.46%
3	0	0	0	156	0	2	158 98.73%
4	0	0	0	0	209	0	209 100.00%
5	0	0	0	0	0	156	156 100.00%
	136 100.00%	137 99.27%	200 100.00%	158 98.73%	219 95.43%	159 98.11%	1009 98.41%

（c）Single-Channel CNN-LSTM

	0	1	2	3	4	5	
0	136	0	0	0	0	0	136 100.00%
1	0	136	0	0	0	0	136 100.00%
2	1	0	206	0	7	0	214 96.26%
3	0	0	0	156	0	2	158 98.73%
4	0	0	0	0	209	0	209 100.00%
5	0	0	0	0	0	156	156 100.00%
	137 99.27%	136 100.00%	206 100.00%	156 100.00%	216 96.76%	158 98.73%	1009 99.01%

（d）Dual-Channel CNN-LSTM

Fig. 9. Confusion Matrices for Each Model.

each model, the true values, and prediction rates for each class. Notably, the dual-channel CNN-LSTM model consistently demonstrates the highest performance across all metrics.

Using the formulas from Table 4, let's calculate the recall, precision, and F1 scores for each model and present the results in Table 5

CNN Model: Under MFCC feature extraction, the CNN model achieved relatively high performance metrics, with an accuracy (ACC) of 96.73 %. Recall, precision, and F1 scores were all around 0.97, indicating sensitivity to spectral information. (See Table 6)

LSTM Model: The LSTM model under MFCC exhibited stable performance, with an ACC of 95.64 %. Recall, precision, and F1 scores were all around 0.95, highlighting its effectiveness in capturing temporal information.

Table 5
Calculation Methods for Evaluation Metrics.

	Calculation Formula	Meaning
ACC	$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}$	The proportion of correctly classified samples by the model out of the total samples
Precision	$\text{Precision} = \frac{TP}{TP + FP}$	For predicted values, the probability of correct predictions among the samples predicted as normal
Recall	$\text{Recall} = \frac{TP}{TP + FP}$	For actual values, the proportion of predicted normals among the samples that are actually normal.

Table 6
Evaluation Metrics for Each Model.

	ACC	Precision	Recall	F1
CNN	96.73 %	96.91 %	97.10 %	0.9700
LSTM	95.64 %	96.13 %	95.68 %	0.9590
Single-Channel CNN-LSTM	98.41 %	98.69 %	98.59 %	0.9864
Dual-Channel CNN-LSTM	99.01 %	99.17 %	99.13 %	0.9915

Single-channel CNN-LSTM Model: This model reached high performance levels with an ACC of 98.41 %, and Recall, precision, and F1 scores all exceeding 0.98, demonstrating the synergistic effect of combining convolutional and long short-term memory network structures under MFCC feature extraction.

Dual-channel CNN-LSTM Model: Achieving the highest performance under MFCC, the dual-channel CNN-LSTM model had an ACC of 99.01 %, with Recall, precision, and F1 scores all surpassing 0.99, showcasing the outstanding performance of combining deep convolution and long short-term memory networks on spectral features.

On the same test set, compared to the other three algorithms, the dual-channel CNN-LSTM model exhibited superior performance: (1) The accuracy of the dual-channel CNN-LSTM model reached 99.01 %, indicating its ability to accurately classify samples of lung sound data. (2) The model's recall was significantly higher than the individual CNN and LSTM models, reaching 99.13 %. High recall means the model can effectively capture positive samples, reducing the risk of false negatives, which is crucial for lung sound classification. (3) The F1 score reached 0.9915, further confirming the model's ability to classify different categories. The F1 score is an indicator that considers both precision and recall, and a high F1 score indicates excellent performance in sample classification.

Therefore, the proposed dual-channel CNN-LSTM model demonstrated outstanding performance in classifying lung sound data. It not only achieved high accuracy but also effectively captured features of different categories, showcasing strong classification capabilities. This model may have broad applications in the field of lung sound classification, potentially improving diagnostic accuracy and efficiency.

5. Conclusion

5.1. Research summary

This study addresses the problem of lung sound signal classification by proposing a deep learning classification model based on dual-channel CNN-LSTM. Firstly, to tackle the issue of uneven dataset distribution, data augmentation techniques were employed to effectively alleviate overfitting and data imbalance during model training. Secondly, the dataset was preprocessed and feature-extracted using Mel-frequency cepstral coefficients (MFCC), transforming lung sound signals into Mel spectrograms as input for the model. Finally, leveraging the advantages of CNN for spatial feature extraction and LSTM for temporal feature extraction, the research carefully tuned parameters and employed K-fold cross-validation to select the optimal model. The test results demonstrate outstanding performance in accuracy and F1 score, reaching 99.01 % and 0.99, respectively. Compared to CNN, LSTM, and the single-channel CNN-LSTM model, the proposed dual-channel CNN-LSTM classification model exhibited higher accuracy and more precise classification.

Experiments indicate that the dual-channel CNN-LSTM lung sound classification model performs well in lung sound classification technology. The application of deep learning in lung sound classification involves leveraging computer simulations of human brain learning and information processing. It analyzes and processes lung sounds through multi-layered neural networks, achieving classification of lung sounds. The application of deep learning in lung sound classification offers four advantages:

Automation: Deep learning enables automatic processing and classification of lung sound signals, reducing the workload for doctors and improving the efficiency and accuracy of medical work.

Big Data Processing: Deep learning requires large amounts of data for model training, and lung sound datasets are typically substantial. Deep learning effectively handles this big data, enhancing the accuracy of lung sound classification.

Efficiency: Deep learning possesses high processing speed and accuracy, enabling rapid classification of a large number of lung sound signals in a short period.

Scalability: Deep learning exhibits good scalability, allowing continuous training and optimization of models to adapt to different requirements for lung sound signal classification.

5.2. Future prospects

While deep learning has shown numerous advantages in lung sound classification, several challenges persist, including but not limited to the following two points: (1) Effectively handling lung sound signals from different patients, respiratory states, and noise environments. (2) Addressing the issue of insufficient datasets. To tackle these challenges, further research is needed to enhance the intelligence of lung sound classification algorithms and elevate the application level of deep learning in this domain.

In the past, home healthcare primarily included basic auxiliary devices such as thermometers. However, the proliferation of electronic medical devices like blood pressure monitors, glucometers, and pulse oximeters has become commonplace. With the advancement of artificial intelligence, more intelligent healthcare devices, such as smart stethoscopes, are expected to enter households in the future. The development of deep learning technology accelerates the integration of lung sound classification into clinical assistance, and the future prospects of lung sound classification technology are broad, encompassing the following three aspects:

(1) Multimodal Data Fusion: In future research, exploring the fusion of lung sound data with other physiological signals (such as electrocardiographic signals and pulse signals) can be attempted to construct multimodal data-driven lung sound classification models. This approach can enhance the robustness and classification performance of the models.
(2) Non-Invasive Sensors and Real-time Monitoring: Traditional lung sound acquisition often requires specific equipment or sensors and is typically conducted in restricted laboratory environments. However, with advancements in non-invasive sensing technologies, such as wearable devices and smartphones, future research can explore how to achieve real-time monitoring and classification of lung sounds based on these devices. This would make lung sound classification technology more convenient and widespread, aiding in the early detection of pulmonary diseases and monitoring changes in health conditions.
(3) Long-term Monitoring and Dynamic Classification: Lung sound classification is typically based on a single auscultation, but the characteristics of certain pulmonary diseases may change over different time periods. Therefore, future research can focus on methods for long-term monitoring and dynamic classification. Continuous collection and analysis of lung sounds can provide a more accurate assessment of disease progression and treatment efficacy.

The future development prospects for lung sound classification research based on deep learning are vast. By exploring research directions such as non-invasive and portable data collection technologies, multimodal data fusion, long-term monitoring, and dynamic classification, we can anticipate further improvements in the performance and applicability of lung sound classification, providing better support for the diagnosis and treatment of pulmonary diseases.

Funding

This research was funded by Natural Science Foundation of China (61901515, 62361038), Natural Science Foundation of Gansu Province (22JR5RA002) and the Innovation Fund for University Teachers of Department of Education of Gansu Province (2023A-019).

CRediT authorship contribution statement

Yipeng Zhang: Software, Writing – original draft, Writing – review & editing. **Qiong Huang:** Supervision. **Wenhui Sun:** Visualization. **Fenlan Chen:** Formal analysis. **Dongmei Lin:** Supervision, Validation. **Fuming Chen:** Methodology, Project administration, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.bspc.2024.106257.

References

[1] The Chronic Non communicable Disease Prevention and Control Center of the China Center for Disease Control and Prevention, and the Statistical Information Center of the National Health Commission Chinese Cause of Death Monitoring Dataset 2019 [M] Beijing: China Science and Technology Press, 2020.

[2] Hu. Jianping, K. Rao, Juncheng qian, etc a study on the economic burden of chronic non communicable diseases in China, Prevention and Control of Chronic Diseases in China 15 (3) (2007) 189–193.

[3] A. Hollman, An ear to the chest: an illustrated history of the evolution of the Stethoscope, J. Royal Soc. Med. 95 (12) (2002) 626–627.

[4] Yilu Ao.The development of a portable electronic heart lung sound stethoscope [D] Chongqing University, 2016, 2.

[5] T.H. Falk, W.Y. Chan, Modulation filtering for heart and lung sound separation from breath sound recordings, Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2008 (2008) 1859–1862.

[6] F. Ayari, M. Ksouri, A.T. Alouani, Computer based analysis for heart and lung signals separation, Int. Conf. Comp. Med. Appl. (2013) 1–6.

[7] I. Mazic, M. Bonkovic, B. Dzaja, Two-level coarse-to-fine classification algorithm for asthma wheezing recognition in children's respiratory sounds, Biomed. Signal Proc. Control 21 (2015) 105–118.

[8] C.H. Chen, W.T. Huang, T.H. Tan, et al., Using k-nearest neighbor classification to diagnose abnormal lung sounds, Sensors 15 (6) (2015) 13132–13158.

[9] N. Sengupta, M. Sahidullah, G. Saha, Lung sound classification using cepstral-based statistical features, Comput. Biol. Med. 75 (2016) 118–129.

[10] X. Zhang, Research on lung sound recognition and diagnosis based on BP neural network, Elect. Test. 13 (2016) 111–113.

[11] X. Zhang, Research and design of auscultation signal recognition system based on deep learning, Jiangsu University (2022), https://doi.org/10.27170/dcnki.gjsuu.2022.000413.

[12] Y. Choi, H. Lee, Interpretation of lung disease classification with light attention connected module, Biomed. Signal Proc. Control 84 (2023) 104695, https://doi.org/10.1016/j.bspc.2023.104695 (Epub 2023 Mar 2. PMID: 36879856; PMCID: PMC9978539).

[13] G. Petmezas, G.A. Cheimariotis, L. Stefanopoulos, B. Rocha, R.P. Paiva, A. K. Katsaggelos, N. Maglaveras, Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function, Sensors (Basel). 22 (3) (2022) 1232, https://doi.org/10.3390/s22031232 (PMID: 35161977; PMCID: PMC8838187).

[14] G. Altan, Y. Kutlu, N. Allahverdi, Deep learning on computerized analysis of chronic obstructive pulmonary disease, IEEE J. Biomed. Health Inform. 24 (5) (2020) 1344–1350.

[15] L. Cui, C. Cui, Z. Liu, et al., Speech emotion recognition using improved MFCC and parallel hybrid models, Comp. Sci. 50 (S1) (2023) 166–172.

[16] D. Liu, H. Yang, W. Hou, B. Wang, A novel underwater acoustic Target recognition method based on MFCC and RACNN, Sensors (Basel) 24 (1) (2024) 273, https://doi.org/10.3390/s24010273 (PMID: 38203134; PMCID: PMC10781205).

[17] Methods In Medicine CAM. Retracted: An Analytical Study of Speech Pathology Detection Based on MFCC and Deep Neural Networks. Comput Math Methods Med. 2023 Dec 13;2023:9829813. doi: 10.1155/2023/9829813. PMID: 38124984; PMCID: PMC10732986.

[18] A. VIJAYAN, B.M. MATHAI, K. VALSALAN, Throatmicrophone speech recognition using MFCC 2017, Int. Conf. Net. Adv. Comput. Techno. (NetACT) (2017) 392–395.

[19] W. Fan, L. Yanxia, L. Liming, et al., Research on sound diagnosis method for power Transformer faults based on deep Learning models, Electr. Technol. 44 (1) (2020) 76–80.

[20] N. Ketkar, J. Moolayil, Deep Learning with.python, Apress, Berkeley, USA, 2021, pp. 197–242.

[21] H. Isik, S. Tasdemir, Y.S. Taspinar, R. Kursun, I. Cinar, A. Yasar, E.T. Yasin, M. Koklu, Maize seeds forecasting with hybrid directional and bi-directional long short-term memory models, Food Sci Nutr. 12 (2) (2023) 786–803, https://doi.org/10.1002/fsn3.3783 (PMID: 38370035; PMCID: PMC10867492).

[22] Hu. Kai, F. Zheng, Lu. Feiyu, et al., A review of behavior recognition algorithms based on deep learning, J. Nanjing Univer. Inform. Technol. (Natural Science Edition) 13 (6) (2021) 730–743.

[23] D. Lihong, X. Chunlang, Y. Ou, et al., A stealing electricity detection method based on CAEs LSTM fusion model, Power Syst. Protect. Cont. 50 (21) (2022) 118–127.

[24] B. Li, J. Wang, Shuiying Liang, et al real time control strategy for AGC based on long short-term memory recurrent neural network, Power Autom. Equi. 42 (3) (2022) 128–134.

[25] M. Huang, T. Wang, Zhinong wei, et al UKF dynamic Harmonic state estimation based on long short term memory networks, Power Syst. Protect. Control 50 (11) (2022) 1–11.

[26] Yu. Mao, H. Shang, Yu. Zhuoqi, Fast identification of synchronous generator groups in power grids based on long short-term memory networks, J. Electr. Eng. 17 (2) (2022) 201–207.

[27] Y. Pang, Q. He, G. Jiang, et al., Spatio - temporal fu-sion neural network for multi-class fault diagnosis of wind turbines based on SCADA Data, Renew-Able Energy 161 (2020) 510–524.

[28] B.M. Rocha, et al., An open access database for the evaluation of respiratory sound classification algorithms, Physiol. Meas. (2019) 40035001.

[29] D. Chen, J. Lin, X. Yi, et al., Classification of underwater acoustic signals based on wavelet packet time-frequency feature and convolutional neural network, Acoustic Technol. 40 (03) (2021) 336–340.