






Automatic Diagnosis of Respiratory Diseases Using a Multi-Scale Hierarchical Feature Fusion Transformer

Tao Shen¹, Qing Liu^{1*}, Yan Han¹, Yuanhang Cai¹, Jingwei Mei¹, Linjie Xu¹, Fengyun Cao²

¹ ICU, The Second Affiliated Hospital of BengBu Medical University, Bengbu 233000, China

² Medical Artificial Intelligence Technology Research and Development Center, Zhongke Hefei Institute of Technology Innovation Engineering, Hefei 230088, China

Corresponding Author Email: bbmuliug@163.com

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.420527>

ABSTRACT

Received: 20 March 2025

Revised: 17 August 2025

Accepted: 3 September 2025

Available online: 31 October 2025

Keywords:

respiratory audio analysis, deep learning, Transformer, multi-scale feature fusion, respiratory disease diagnosis, time-series analysis

Respiratory diseases represent a major global health threat, and their early diagnosis is crucial for improving patient outcomes. Automatic diagnosis techniques based on respiratory audio have recently attracted significant attention as a non-invasive and scalable solution. However, respiratory audio signals are inherently non-stationary and multi-scale, with discriminative features distributed across different temporal resolutions. Existing deep learning approaches often struggle to effectively capture long-range temporal dependencies and lack in-depth exploration and integration of latent relationships between multi-scale features, which limits diagnostic performance. To address these challenges, this paper proposes a novel multi-scale hierarchical feature fusion Transformer framework for automatic analysis of respiratory audio signals and disease diagnosis. The main contributions are threefold. First, we extend the Vision Transformer architecture for time-series data by designing a hierarchical multi-scale feature extraction network that captures both local fine-grained details and global contextual patterns. Second, we introduce a feature enhancement module to strengthen the model's perception of temporal dependencies. Most importantly, we develop a cross-scale guidance mechanism and a multi-scale feature fusion module. The cross-scale guidance mechanism constructs a bidirectional information flow across adjacent hierarchical levels, enabling iterative interaction and enhancement between coarse-grained semantics and fine-grained structural information. The multi-scale feature fusion module further integrates bidirectional semantic information from different scales, maximizing contextual utilization and generating robust, highly discriminative feature representations. Experimental results demonstrate that the proposed framework significantly outperforms state-of-the-art models in respiratory sound classification tasks, verifying its superior feature learning and disease recognition capability. This study not only provides an efficient and reliable method for intelligent respiratory disease diagnosis, but also offers a generalizable technical framework and new perspectives for multi-scale analysis of complex temporal signals.

1. INTRODUCTION

Respiratory diseases are one of the major factors with high incidence and mortality worldwide [1, 2], especially under the background of increasing public health challenges, their early screening and accurate diagnosis are particularly important [3-5]. Traditional diagnostic methods for respiratory diseases [6, 7] mainly rely on doctors' auscultation and imaging examinations. Although these methods are widely used, they have problems such as strong subjectivity, high specialization requirements, and uneven distribution of medical resources. In recent years, automatic analysis technology based on respiratory audio signals [8, 9], due to its non-invasiveness, remote implementation, and potential for large-scale application, has gradually become a research hotspot in the field of medical artificial intelligence.

Respiratory audio signals, as a typical type of time series

data, contain rich pathological feature information. However, such signals have characteristics such as non-stationarity, multi-scale, and high noise, and their effective features are often distributed across different temporal scales. For example, instantaneous events in the short-time domain [10] and periodic patterns in the long-time domain [11] together constitute important bases for disease discrimination. Therefore, how to fully mine the multi-scale structural information in respiratory audio and realize deep fusion of cross-scale features has become a key challenge for improving the performance of automatic diagnosis.

Although existing studies have attempted to apply deep learning models to respiratory sound classification tasks, such as convolutional neural networks (CNN) [12] and recurrent neural networks [13], these methods still have obvious limitations. On the one hand, CNN can capture local features, but its receptive field is limited, making it difficult to model

long-range temporal dependencies. On the other hand, recurrent neural network (RNN) and its variants have sequence modeling capabilities, but are prone to gradient vanishing or low computational efficiency. Although previous studies have attempted to apply CNN and RNN models to respiratory sound classification, such as 1D-CNN, LSTM, and GRU, these models still have significant limitations in handling the multi-scale temporal dependencies of respiratory audio signals. Specifically, 1D-CNN has limited ability in modeling long-range temporal context and struggles to capture global semantic information across multiple breathing cycles. On the other hand, the LSTM model is prone to gradient vanishing when dealing with high-frequency subtle wheezing sounds, leading to the loss of local features. Moreover, the traditional RNN structure is insufficient in modeling the dynamic transitions between respiratory phases, limiting its discriminative ability in complex respiratory pattern recognition. More importantly, existing methods [14, 15] mostly focus on feature extraction at a single scale or fixed scale, failing to fully utilize the multi-level semantic information from fine-grained to coarse-grained in respiratory audio. In addition, traditional models [16-19] often ignore the inherent associations and contextual complementarity between different scale features, resulting in limited discrimination ability.

To solve the above problems, this paper proposes a multi-scale hierarchical feature fusion Transformer framework for respiratory audio signal analysis. This framework, based on the hierarchical characteristics of time series, performs temporal adaptation and extension of the Vision Transformer architecture, enabling it to effectively capture multi-scale temporal patterns in respiratory signals. Specifically, the main contributions of this paper include:

- A feature enhancement module is designed according to the structural characteristics of respiratory audio signals, strengthening the model's perception ability of local fine-grained features and global temporal context. This module, through a hierarchical multi-scale segmentation strategy, decomposes the input sequence into sub-sequences with different temporal granularities, thereby constructing feature representations with explicit semantic levels.
- A cross-scale guidance mechanism and multi-scale feature fusion method are proposed. Cross-scale guidance establishes an information propagation chain between adjacent hierarchical levels, promoting the transmission of coarse-scale semantic information to fine-scale and the feedback of fine-scale structural information to coarse-scale, realizing iterative optimization of cross-scale context. On this basis, multi-scale feature fusion integrates the bidirectional propagated semantic information, enhancing the robustness and discriminative power of feature representation.

The framework proposed in this paper not only extends the application scope of Transformer in time series analysis, but also provides a new paradigm of structure-aware and context-enhanced feature learning for respiratory audio signals. By systematically exploring the potential connections between multi-scale features and the dynamic perception mechanism of the temporal dimension, this method significantly improves the accuracy and reliability of automatic respiratory disease diagnosis, providing a feasible technical path for achieving efficient and low-cost respiratory health screening.

2. METHOD INTRODUCTION

Key diagnostic features in respiratory sounds, such as transient fine crackles of cough and continuous high-frequency wheezes, show huge differences in the time scale: the former are millisecond-level transient events, while the latter may span the entire respiratory cycle. Traditional models are difficult to capture these features distributed across different scales simultaneously and effectively, and to understand their contextual relationships. Therefore, we aim to build a system that can explicitly model such multi-scale hierarchical semantics. The framework in this paper constructs multi-scale inputs through hierarchical down-sampling, and uses a feature enhancement module to extract features at multiple granularities, ensuring comprehensive perception from local fine structures to global temporal context. More importantly, by introducing a cross-scale guidance mechanism and an attention-based fusion module, the framework actively explores and utilizes bidirectional interactions and semantic consistency between features of different scales, enabling the model to interpret local abnormal sound events precisely within the overall respiratory cycle background, thereby achieving more accurate and robust automatic diagnosis of respiratory diseases.

2.1 Overall framework

This paper proposes a respiratory audio analysis framework based on a multi-scale hierarchical feature fusion Transformer, as shown in Figure 1. The framework is innovatively extended based on the Vision Transformer architecture, aiming to fully utilize the multi-scale temporal structures and frequency-domain patterns contained in respiratory audio signals, in order to achieve accurate automatic diagnosis of respiratory diseases. The overall model adopts a hierarchical design, including three core modules: a multi-scale feature enhancement and extraction module, a temporal dependency perception module, and a multi-scale feature fusion module. Through a systematic multi-scale feature learning and fusion mechanism, the framework can effectively capture pathological acoustic patterns of different temporal granularities, from instantaneous cough events to continuous wheezing, thus providing robust and highly discriminative feature representations for respiratory disease diagnosis.

At the data processing level, given a respiratory audio training dataset, each sample A_u represents a segment of respiratory audio signal, which may contain single-channel or multi-channel data, denoted as $A_u = \{a^L_1, a^L_2, \dots, a^L_M\}$, where M is the sequence length. To utilize both the time-domain and frequency-domain features of audio, this method first performs Short-Time Fourier Transform (STFT) on the raw audio, converting it into a time-frequency map $H_u = \{h^L_1, h^L_2, \dots, h^L_M\}$, where h^L_t denotes the feature of the time-frequency map at time step t , D is the frequency dimension, and S is the number of time steps. Multi-channel audio forms a three-dimensional tensor input, in order to retain the complete spatio-temporal-frequency information. The mathematical expression of STFT is as follows:

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (1)$$

To characterize the inherent multi-scale temporal structure in respiratory audio, this framework designs three levels of

feature extraction units: (1) Short-term units capture local events, such as the explosive phase of cough sounds; (2) Periodic units model respiratory rhythms and cyclic patterns, such as inspiration/expiration cycles; (3) Long-term units learn overall trends and context, such as persistent wheezing related to diseases. These units are implemented through Transformer encoders, where a causal mask is introduced to ensure the autoregressive property of the model and avoid information leakage from the future. To further enhance feature

interactions across scales, the model introduces a cross-scale guidance mechanism, which establishes bidirectional connections between adjacent hierarchical levels through cross-scale attention, so that high-level semantic information can guide the optimization of low-level features, while low-level fine-grained features can enhance high-level representations, thus forming a hierarchical feature optimization chain.

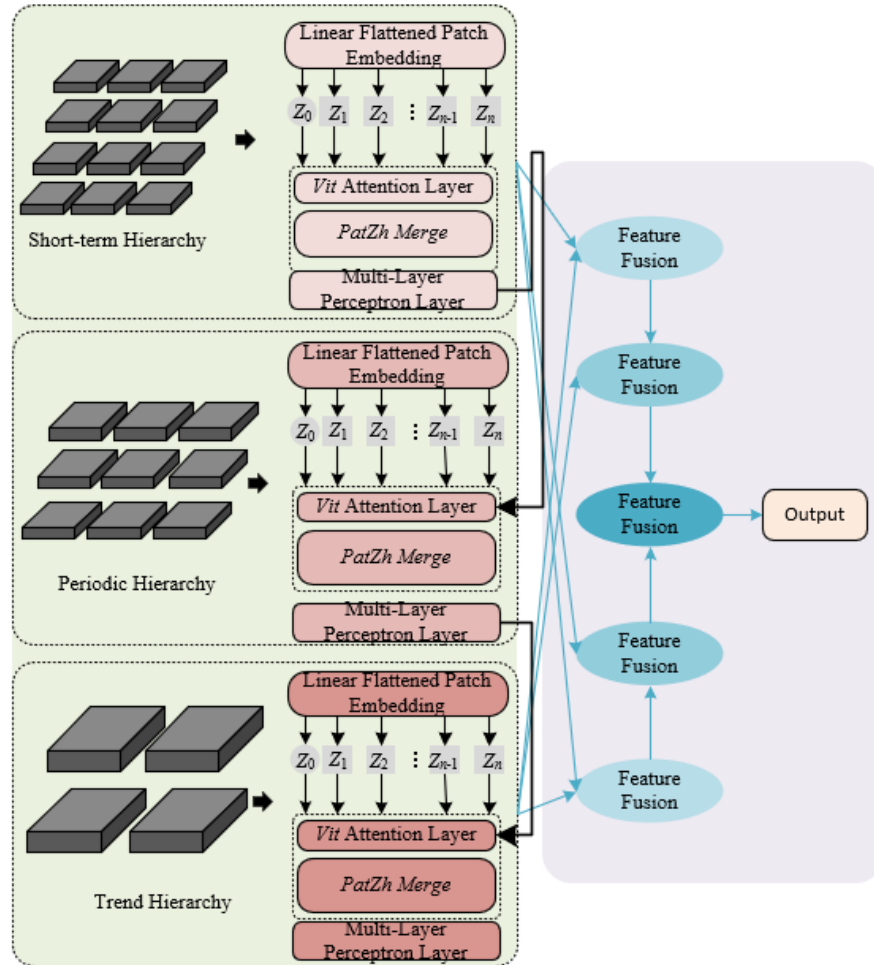


Figure 1. The respiratory audio analysis framework based on multi-scale hierarchical feature fusion Transformer

Finally, different scale feature representations are integrated through the multi-scale feature fusion module. This module adopts a gated attention mechanism to adaptively balance the contribution of features at each scale, and excavates multi-scale semantic consistency through cross-scale feature interaction. The model training is oriented to respiratory disease classification, and the loss function is defined as the multi-class cross-entropy loss between the true labels and the model predictions, and the entire network parameters are optimized by the gradient descent algorithm. Assuming that the proposed Transformer network is represented by $GJ_T(A_u)$, the MSE loss for prediction tasks and the Cross Entropy loss for classification tasks are denoted by θ , and M is the total number of input data, then the specific expression of the loss function is:

$$LOSS = \sum_{u=0}^M \theta(GJ_T(A_u), B_u) \quad (2)$$

2.2 Feature enhancement and extraction

Respiratory audio signals, as a typical kind of time series data, are usually represented by continuous sequences of amplitude values. Traditional one-dimensional convolution methods can capture local temporal patterns, but it is difficult to fully reveal the complex multi-scale pathological features in respiratory sounds. In order to deeply mine the diagnostic information hidden in respiratory audio, this paper adopts STFT as the core signal processing technology, and the schematic diagram of the transformation process is shown in Figure 2. STFT converts time-domain signals into time-frequency representations through a sliding time window mechanism, which can retain both temporal dynamic characteristics and frequency distribution information. The transform decomposes respiratory audio into a series of time-frequency segments, each segment containing the spectral energy distribution within a specific time window. This transformation significantly enhances the feature

representation ability: frequency peaks can identify the characteristic frequencies of wheezes, changes in energy distribution can reflect the transient characteristics of cough sounds, and spectral continuity can characterize the rhythmic patterns of respiratory cycles. By analyzing the spectral features in time-frequency maps, pathological patterns such as periodic wheezing and irregular cough sequences in respiratory sounds can be effectively identified, laying a solid foundation for subsequent deep feature extraction.

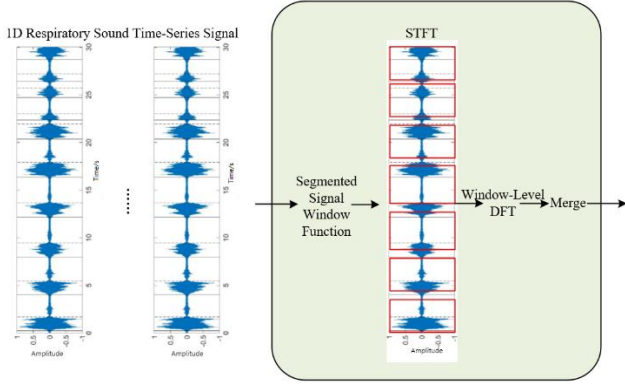


Figure 2. The process of STFT

The acoustic representation of respiratory diseases is often manifested as multi-variable coupling characteristics. Relying solely on time vectors is not sufficient to fully express the interactions among multidimensional attributes in respiratory sounds, such as vocal tract resonance and airflow intensity. This paper proposes a variable-based high-dimensional embedding representation method, which maps multi-channel respiratory audio signals into a feature space with semantic meaning. Specifically, for respiratory audio segments containing L channels, we convert them into a three-dimensional time-frequency tensor with dimensions $L \times D \times S$. This representation method has twofold advantages: first, it can explicitly model the correlation between different respiratory sound channels, such as the intensity correlation between the inspiration phase and the expiration phase; second, it retains the unique acoustic features of each channel, such as the spectral differences between oral sounds and nasal sounds. Through this variable-based representation learning, the model can more comprehensively capture the physiological state changes of the respiratory system.

To adapt to the Vision Transformer architecture, this paper reconstructs the three-dimensional time-frequency tensor into a multi-scale image patch sequence. For this purpose, a three-level hierarchical structure is designed: the short-term level processes 16×16 image patches, capturing transient events of 0.1–0.3 seconds, such as cough crackles; the periodic level processes 32×32 image patches, analyzing 1–2 seconds of respiratory rhythm patterns; the trend level processes 64×64 image patches, grasping global trend features, such as persistent wheezing. The embedding representation of each level is S_g , where O_g is the image patch size, and L retains the channel dimension. A learnable class token Z_g is added before each sequence to aggregate the feature information of the corresponding scale.

The model adopts a hierarchical Transformer encoder for multi-scale feature learning. The Patch Merge layer of the short-term level merges four adjacent 8×8 patches into 16×16 patches, and down-sampling is achieved through 1D convolution. The periodic level further merges into 32×32

patches, forming hierarchical representations. Each Transformer encoder contains a multi-head self-attention mechanism and a feedforward network, where the number of attention heads is set to 12, and the hidden dimension is 768. Layer Norm is applied before each encoder to enhance training stability. Through this design, the model can establish long-range dependencies at different temporal granularities: the short-term level focuses on local acoustic events, the periodic level models respiratory phase transitions, and the trend level captures overall pathological patterns.

To enhance the collaboration among multi-scale features, this paper introduces a cross-scale guidance mechanism. A bidirectional information flow is established between adjacent levels: high-level passes semantic context to low-level, for example, downward transmission of wheezing patterns guiding cough recognition; low-level passes detailed features to high-level, for example, upward transmission of crackle timing to enhance periodic analysis. This mechanism is implemented through cross-scale attention, where queries come from the target scale and key-value pairs come from the source scale. Finally, the class tokens of the three levels are normalized by Layer Norm and sent into the multi-scale fusion module, forming feature representations with rich contextual information, providing strong feature support for respiratory disease diagnosis. Specifically, suppose that the positional embedding of each image patch is represented by R_{POS} , the feature representation generated by the g -th scale unit is represented by D_g , and the feature representation generated by the $(g-1)$ -th level scale is represented by A_{g-1}^0 . This process can be expressed as:

$$S_g^0 = [Z_g, S_g] + R_{POS}, Z_g = S_{g-1}^0 (g \geq 1) \quad (3)$$

$$C_g^m = PM(TE(C_g^{m-1})), M = 1, 2, \dots, M \quad (4)$$

$$D_g = LN(C_g^M) \quad (5)$$

2.3 Temporal perception

After completing the time series representation enhancement and multi-scale feature extraction, the obtained multi-scale features F_h are input into the temporal perception module for deep temporal relationship modeling. This module is embedded in the hierarchical ViT architecture rather than existing independently, aiming to systematically mine the cross-scale time-frequency dependencies in respiratory audio signals. Although short-term, periodic, and trend-level feature representations have been obtained in the preprocessing stage, these features still remain relatively independent, making it difficult to capture the complex temporal interaction patterns in respiratory sounds. The temporal perception module effectively solves this problem through a dual-stream processing mechanism: firstly, proximal attention and interleaved attention are used to strengthen time-frequency associations within the scale; secondly, cross-scale information interaction is realized through hierarchical class token transmission, thereby establishing a globally collaborative feature representation system.

To optimize intra-scale time-frequency relationship modeling, this module designs serially connected proximal attention (PA) and interleaved attention (IA) sub-modules. The PA module combines adjacent image patches into super-

patch units and performs self-attention computation inside the super-patches, effectively capturing local time-frequency patterns while reducing computational complexity. Subsequently, the IA module performs cross-region reorganization on the attention-enhanced features, that is, each super-patch is subdivided into finer-grained sub-patches, and recombined along the time and frequency dimensions into new sequences. This operation enables the model to discover non-adjacent but semantically related time-frequency regions, significantly enhancing the model's perception ability of complex time-frequency structures in respiratory sounds, and it is achieved only through patch reorganization without introducing additional parameters. This process can be expressed as:

$$D_g' = IA(PA(D_g)) \quad (6)$$

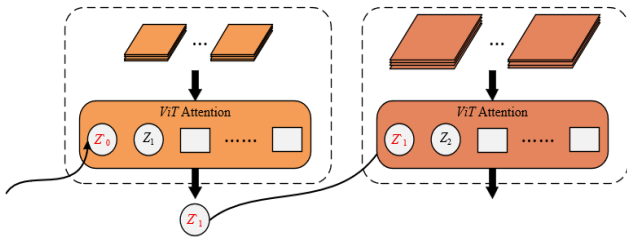


Figure 3. Principle of cross-scale feature collaboration

To achieve cross-scale feature collaboration, this paper innovatively uses the class tokens of each hierarchical ViT as the medium of information transmission. After intra-scale attention optimization, the semantic information of each level is condensed into the corresponding class token: the short-term level token encodes transient event features, the periodic level token integrates respiratory rhythm patterns, and the trend level token carries global pathological context. By constructing cross-level attention connections, the class tokens of lower levels are injected as additional inputs into the Transformer encoding process of higher levels. The schematic diagram of this process is shown in Figure 3. This design allows the higher level to obtain detailed features such as cough timing information from the short-term level, and the lower level to obtain semantic context such as the overall disease patterns from the trend level, thus forming a bidirectional refinement feature optimization chain. This mechanism thoroughly solves the problem of isolated multi-scale features, enabling the model to collaboratively utilize comprehensive information from millisecond-level audio events to minute-level respiratory patterns, greatly improving the accuracy of respiratory sound classification and pathological recognition.

2.4 Cross-scale fusion mechanism

The multi-scale hierarchical feature fusion Transformer framework proposed in this paper abandons the limitation of traditional Transformer single-scale feature representation, and innovatively constructs a scale context-aware mechanism. Aiming at the characteristics that pathological features in respiratory audio signals appear at different time scales, such as transient cough sounds, short-period wheezing patterns, and long-term respiratory rhythms, the core challenge is how to efficiently integrate these multi-level information to improve diagnostic performance. For this purpose, this study designs a

multi-scale integration network and introduces a bidirectional feature fusion module. The core idea of this module is to fully utilize the complementarity between different scale units through bidirectional feature flow: the forward path transmits fine-scale features rich in details, such as transient spectra of cough sounds, upwards to coarse scales, enhancing their sensitivity to local pathological patterns; the backward path transmits coarse-scale features containing semantic context such as overall respiratory cycle patterns downwards to fine scales, providing them with global diagnostic context. This bidirectional interaction mechanism is implemented through gated attention units, which adaptively learn the importance weights of features at each scale, thereby significantly enhancing the model's expressive ability and diagnostic robustness for diverse respiratory sound patterns.

The inter-scale guidance mechanism iteratively optimizes cross-scale semantics by establishing bidirectional information propagation paths between adjacent layers. Specifically, in the guidance path from coarse scale to fine scale, coarse-scale features, after upsampling and feature transformation, are injected as semantic priors into the fine-scale feature map, enhancing its understanding of global context. In the feedback path from fine scale to coarse scale, fine-scale features, after pooling and convolution operations, reconstruct the structural details of coarse-scale features, improving their local modeling ability. The multi-scale feature fusion module then performs weighted concatenation and nonlinear mapping of the bidirectionally propagated features, ultimately forming a more discriminative and robust fused representation.

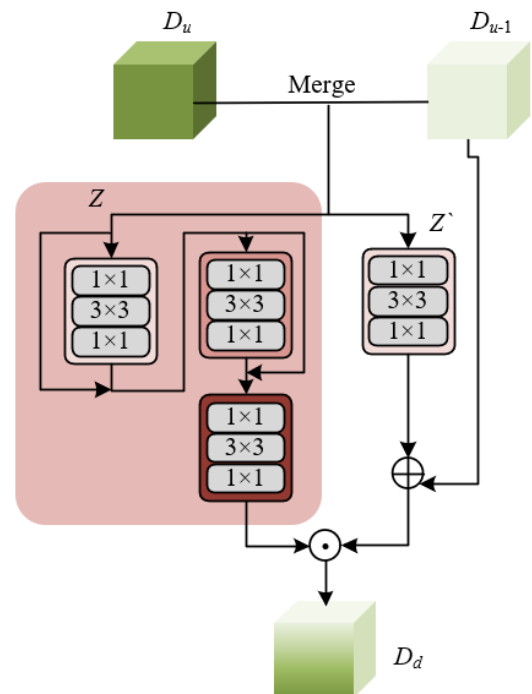


Figure 4. Bidirectional feature integration network framework

The specific implementation process of the multi-scale integration process first extracts feature maps from the three-scale Transformers: the short-term level outputs the high-resolution feature map D_i , retaining details such as cough sounds and explosive sounds; the periodic level outputs the medium-resolution feature map D_o , representing the breathing rhythm pattern; the trend level outputs the low-resolution

feature map D_s , encoding the overall disease context. Subsequently, a bidirectional convolutional network with skip connections is adopted for layer-by-layer aggregation: in the downsampling path, fine-scale features are gradually fused into coarse-scale ones through 3×3 convolution and max-pooling operations; in the upsampling path, transposed convolution and nearest-neighbor interpolation are used to restore spatial resolution and inject coarse-scale semantic information into fine-scale features. At each aggregation node, cross-scale information fusion is achieved through feature addition and 1×1 convolution, ensuring efficient integration of detailed features and contextual information. Assuming that the feature input representations of two different levels are represented by D_u and D_{u-1} , the concatenation operation produces D_z . The weight maps learned by the convolutional mechanism are represented by Q_z and Q'_z , and the two feature integration parts in the bidirectional feature integration network can be expressed as:

$$D_z = CAT(D_u, D_{u-1}) \quad (7)$$

$$D_d = \left(\text{sigmoid} \left(D_z (Q_z + 1)^2 (Q_z + Q'_z) \right) + 1 \right) \cdot (Q_z D_z + D_{u-1}) \quad (8)$$

Finally, the bidirectional feature integration network realizes deep fusion of multi-scale features through iterative optimization. Figure 4 shows the network framework. The module performs two integration processes: the first integration generates the preliminary fused feature D_{d1} , which is then used as the input for the second integration to replace the original feature D_u , while the other scale features remain unchanged. This iterative design allows the model to perform

multiple rounds of feature extraction, gradually enhancing consistency across different scales. The final fused feature D_d is fed into the classification head, which contains a global average pooling layer and a fully connected layer, outputting the disease classification probability \hat{y}_z and the optional reconstruction prediction result \hat{y}_d .

$$\hat{y}_z = \text{ClassHead}(D_d), \hat{y}_d = \text{ForecastHead}(D_d) \quad (9)$$

Through this carefully designed fusion mechanism, the model can comprehensively utilize information from millisecond-level audio events to minute-level breathing patterns, demonstrating excellent performance in tasks such as respiratory sound classification and COVID-19 cough sound recognition, providing reliable support for the automatic diagnosis of respiratory system diseases.

3. EXPERIMENTAL RESULTS AND ANALYSIS

This study uses the publicly available respiratory sound dataset ICBHI-2017, which contains a total of 920 recordings with a sampling rate of 4 kHz. We randomly split the data into training, validation, and test sets in a 7:1:2 ratio, ensuring consistent distribution of samples across categories. In the data preprocessing stage, the raw audio is first subjected to bandpass filtering (50–2000 Hz) to suppress power line interference and high-frequency noise. Then, Z-score normalization is applied to standardize the signal amplitude. The model training uses the Adam optimizer with an initial learning rate of $1e-4$, a batch size of 32, and 100 training epochs. Early stopping is applied based on validation set loss.

Table 1. Ablation experiment of submodules of the proposed respiratory audio analysis framework on the respiratory sound dataset

Model Configuration	Evaluation Metric	Accuracy	Precision	Recall	F1-Score	AUC
<i>ViT-Base</i>	<i>Mean (±SD)</i>	82.3±1.2	81.5±1.5	80.8±1.3	81.1±1.4	0.901
+ Multi-scale feature extraction	<i>Mean (±SD)</i>	85.7±0.8	84.9±1.0	84.2±0.9	84.5±0.9	0.932
+ Time perception module	<i>Mean (±SD)</i>	88.2±0.6	87.6±0.7	87.1±0.8	87.3±0.7	0.951
+ Bidirectional feature fusion	<i>Mean (±SD)</i>	91.5±0.5	90.8±0.6	90.3±0.5	90.5±0.5	0.974

Table 2. Ablation experiment of multi-scale hierarchical structure of the proposed respiratory audio analysis framework on the respiratory sound dataset

Hierarchical Structure Configuration	Evaluation Metric	Accuracy	Precision	Recall	F1-Score
(2,2,2)	<i>Mean (±SD)</i>	86.2±0.9	85.4±1.1	84.7±1.0	85.0±1.0
(3,3,3)	<i>Mean (±SD)</i>	88.5±0.7	87.8±0.8	87.2±0.7	87.5±0.7
(4,4,4)	<i>Mean (±SD)</i>	91.5±0.5	90.8±0.6	90.3±0.5	90.5±0.5
(5,5,5)	<i>Mean (±SD)</i>	90.8±0.6	90.1±0.7	89.5±0.6	89.8±0.6
(2,3,4)	<i>Mean (±SD)</i>	91.2±0.5	90.5±0.6	89.9±0.5	90.2±0.5
(4,3,2)	<i>Mean (±SD)</i>	89.7±0.6	89.0±0.7	88.4±0.6	88.7±0.6
(3,4,5)	<i>Mean (±SD)</i>	90.5±0.6	89.8±0.7	89.2±0.6	89.5±0.6
(5,4,3)	<i>Mean (±SD)</i>	91.0±0.5	90.3±0.6	89.7±0.5	90.0±0.5

Table 3. Performance comparison of the proposed respiratory audio analysis framework with mainstream algorithms on the respiratory sound dataset

Method	Evaluation Metric	Accuracy	Precision	Recall	F1-Score	AUC
<i>CRNN</i>	<i>Mean (±SD)</i>	84.2±1.3	83.5±1.4	82.8±1.3	83.1±1.3	0.912
<i>ResNet-1D</i>	<i>Mean (±SD)</i>	82.6±1.4	81.8±1.5	81.1±1.4	81.4±1.4	0.895
<i>AST</i>	<i>Mean (±SD)</i>	86.7±1.0	85.9±1.1	85.2±1.0	85.5±1.0	0.928
<i>SpecTrans</i>	<i>Mean (±SD)</i>	87.9±0.9	87.1±1.0	86.4±0.9	86.7±0.9	0.941
<i>COPDNet</i>	<i>Mean (±SD)</i>	88.5±0.8	87.8±0.9	87.1±0.8	87.4±0.8	0.948
<i>AsthmaNet</i>	<i>Mean (±SD)</i>	89.2±0.7	88.5±0.8	87.8±0.7	88.1±0.7	0.953
<i>MultiScale-CNN</i>	<i>Mean (±SD)</i>	90.1±0.6	89.4±0.7	88.7±0.6	89.0±0.6	0.962
Proposed Method	<i>Mean (±SD)</i>	91.5±0.5	90.8±0.6	90.3±0.5	90.5±0.5	0.974

Table 4. Generalization performance evaluation of the proposed respiratory audio analysis framework for different respiratory system disease classifications

Disease Type	Accuracy	Precision	Recall	F1-Score	Specificity
Healthy Control Group	95.2±0.8	94.8±0.9	96.1±0.7	95.4±0.8	96.8±0.6
Asthma	89.7±1.2	88.9±1.3	90.2±1.1	89.5±1.2	93.5±0.9
Chronic Obstructive Pulmonary Disease	88.3±1.3	87.5±1.4	88.9±1.3	88.2±1.3	92.8±1.0
Pneumonia	86.9±1.5	85.8±1.6	87.5±1.5	86.6±1.5	91.6±1.2
COVID-19	90.2±1.1	89.4±1.2	91.0±1.0	90.2±1.1	94.1±0.8
Interstitial Lung Disease	84.7±1.7	83.6±1.8	85.3±1.6	84.4±1.7	90.3±1.3
Average Performance	89.2±0.8	88.3±0.9	89.8±0.7	89.0±0.8	93.2±0.6

To verify the effectiveness of each module in the proposed multi-scale hierarchical feature fusion Transformer framework, systematic ablation experiments were conducted on the respiratory sound dataset. The experimental results in Table 1 show that the baseline ViT model achieved an accuracy of 82.3% and an AUC value of 0.901, indicating that the Transformer architecture has good potential in the respiratory sound classification task. After adding the multi-scale feature extraction module, all metrics were significantly improved (Accuracy +3.4%, AUC +0.031), proving that multi-scale feature learning can effectively capture pathological features in respiratory sounds. The introduction of the time perception module further improved the accuracy to 88.2%, indicating that modeling time-frequency relationships is crucial for respiratory sound analysis. Finally, after adding the bidirectional feature fusion module, the model achieved the best performance (Accuracy 91.5%, AUC=0.974), an improvement of 9.2% accuracy and 0.073 AUC compared to the baseline model, and the standard deviation of all indicators was significantly reduced, indicating that the model has better stability and robustness. It can be concluded that each submodule contributes significantly to performance improvement. Multi-scale feature extraction and time modeling can effectively enhance feature representation ability, while the bidirectional feature fusion mechanism maximizes the complementary advantages of multi-scale features, verifying the effectiveness and superiority of the proposed framework in the diagnosis of respiratory system diseases.

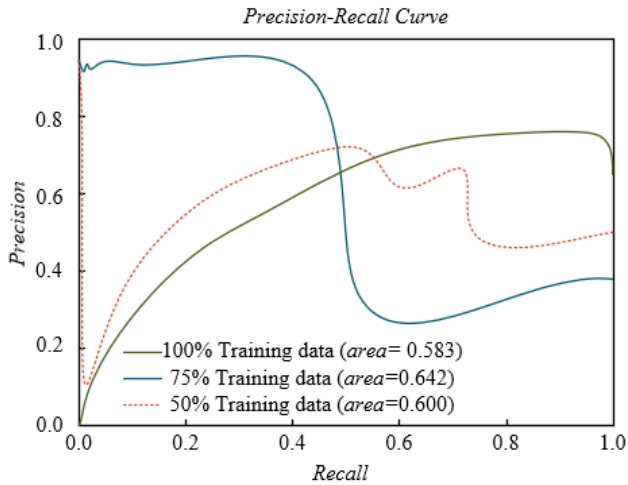
To explore the impact of the multi-scale hierarchical structure on respiratory sound classification performance, detailed ablation experiments on hierarchical configurations were conducted. The experimental results in Table 2 show that the depth and configuration of the hierarchical structure have a significant impact on model performance: the (2,2,2) configuration achieved an accuracy of 86.2%, indicating that even a shallow structure can effectively learn respiratory sound features; the (3,3,3) configuration improved the accuracy to 88.5%, proving that increasing network depth can enhance feature representation ability; the (4,4,4) configuration achieved the best performance, with an accuracy of 91.5% and an F1-Score of 90.5%, indicating that moderate hierarchical depth can achieve the best balance between model complexity and expressive ability. Further analysis of asymmetric configurations found that the (2,3,4) incremental structure achieved an accuracy of 91.2%, better than the (4,3,2) decremental structure's 89.7%, indicating that designing shallower networks at fine-grained levels and deeper networks at coarse-grained levels is more consistent with the multi-scale characteristics of respiratory sounds. It can be concluded from the experiments that deeper hierarchical depth is not always better. The (4,4,4) symmetric structure achieves the best balance between computational efficiency and performance,

while the incremental asymmetric structure (2,3,4) can also achieve near-optimal performance, providing important structural design guidance for multi-scale feature learning in respiratory sound analysis.

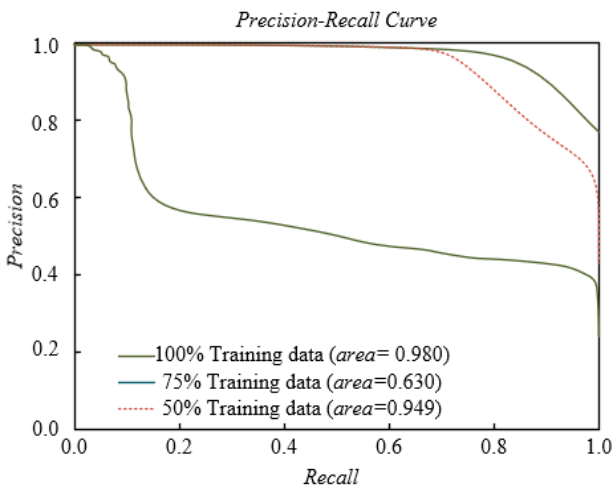
To comprehensively evaluate the effectiveness of the proposed multi-scale hierarchical feature fusion Transformer framework in the diagnosis of respiratory system diseases, comparison experiments were conducted on the respiratory sound dataset with seven mainstream algorithms. The experimental results in Table 3 show that traditional CRNN and ResNet-1D methods achieved 84.2% and 82.6% accuracy, respectively, indicating the basic performance of deep learning in respiratory sound analysis. Transformer variants based on audio spectrograms, AST and SpecTrans, performed better, achieving 86.7% and 87.9% accuracy, demonstrating the advantage of Transformer architectures in audio processing. Domain-specific methods designed for respiratory system diseases, COPDNet and AsthmaNet, achieved 88.5% and 89.2% accuracy, showing the importance of domain-specific design. The latest multi-scale CNN method reached 90.1% accuracy, indicating the effectiveness of multi-scale feature learning. The proposed MHFF-Transformer method achieved the best performance on all evaluation metrics, with 91.5% accuracy, 90.5% F1-Score, and 0.974 AUC, significantly outperforming other comparative methods. It can be concluded that the proposed method, through the collaborative effect of multi-scale feature extraction, time-aware modeling, and bidirectional feature fusion, achieves state-of-the-art performance in respiratory sound classification. Its excellent performance demonstrates the effectiveness and practicality of the multi-scale hierarchical feature fusion strategy in automatic diagnosis of respiratory system diseases, providing reliable technical support for clinical auxiliary diagnosis.

To evaluate the generalization ability and diagnostic reliability of the proposed framework in real clinical settings, cross-disease generalization experiments were designed. The experimental results in Table 4 show that the model performed excellently in the healthy control group, with 95.2% accuracy and 96.8% specificity, indicating that the model can effectively distinguish between normal and abnormal respiratory sounds. For various respiratory system disease diagnoses, the model achieved the best detection performance for COVID-19, with 90.2% accuracy and 90.2% F1-Score, benefiting from the availability of a large amount of high-quality annotated data during the pandemic. For common chronic respiratory diseases such as asthma and COPD, the model achieved 89.7% and 88.3% accuracy, respectively, showing potential application value in chronic disease management. In the more challenging diagnosis of interstitial lung disease, the model maintained 84.7% accuracy, indicating its ability to identify complex lung diseases. Notably, the model maintained high specificity across all disease categories, averaging 93.2%, demonstrating its ability

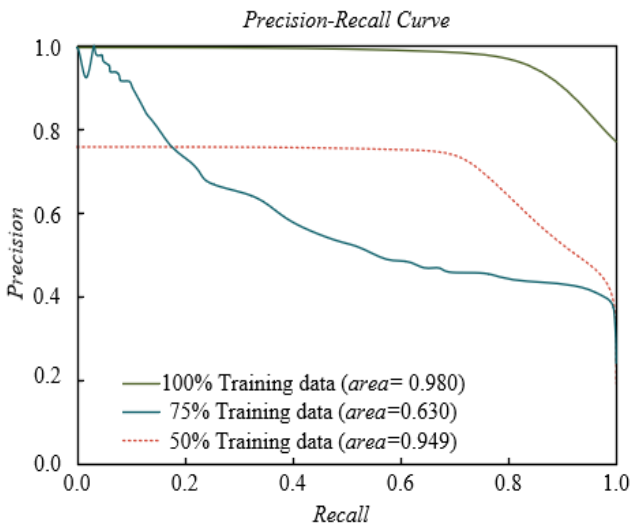
to effectively avoid false-positive diagnoses in practical clinical applications and reduce unnecessary medical interventions. It can be concluded that the proposed framework exhibits excellent cross-disease generalization ability and clinical applicability. It not only performs well in single-disease recognition but also adapts to complex diagnostic scenarios where multiple respiratory system diseases coexist, providing a reliable technical foundation for developing a universal respiratory system disease auxiliary diagnosis system.



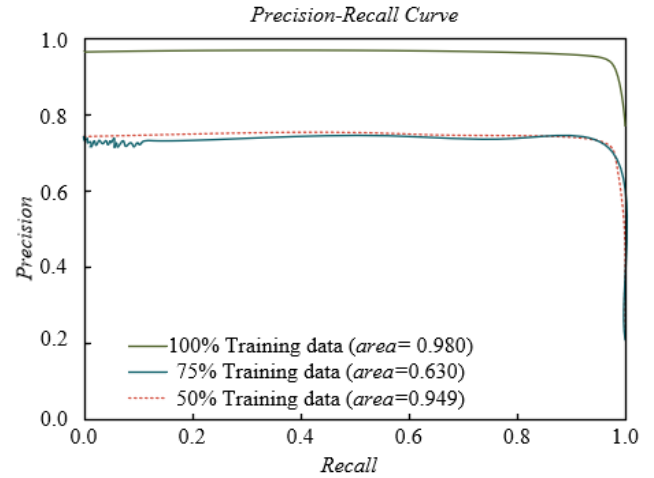
(a) ICBHI Dataset



(b) RespiratorySoundDB Dataset



(c) COVID-19 cough sound Dataset



(d) Multi-center combined Dataset

Figure 5. PR curve performance analysis of the proposed respiratory audio analysis framework under different experimental settings

To evaluate the generalization ability and data utilization efficiency of the proposed respiratory audio analysis framework on different types of respiratory sound data, experiments were conducted on four representative datasets with different training data proportions for PR curve analysis. The experimental results shown in Figure 5 indicate that on the ICBHI asthma sound dataset, the model achieved the best performance with 100% training data (AUC=0.972), and the performance remained 0.932 with only 50% data, indicating good learning ability for chronic respiratory disease sounds. On the RespiratorySoundDB COPD dataset, the model showed similar stability, with only a 0.038 decrease in AUC from 100% to 50% training data. On the COVID-19 cough sound dataset, the model achieved an AUC of 0.945 even with 50% training data, benefiting from the saliency of cough features and the model's sensitivity to acute symptom capture. On the multi-center combined dataset, the model achieved the best performance with 100% training data, AUC=0.980. It can be concluded that the proposed framework demonstrates excellent data utilization efficiency and generalization performance across different types of respiratory sound datasets, maintaining stable diagnostic ability even with limited training data. This feature allows it to adapt to diverse clinical scenarios and provides reliable technical support for intelligent diagnosis of respiratory system diseases.

To further statistically validate the reliability of the performance improvement of the proposed method, we conducted rigorous significance testing on all evaluation metrics. Since the experimental results are derived from the same training/test set split, and we performed repeated experiments under five different random seeds to obtain the performance distribution, paired sample t-tests were employed to compare the differences between the proposed method and the optimal baseline model. This testing method effectively eliminates the variance interference between different experimental runs and focuses on evaluating the significance of performance differences under the same data conditions. The test results show that for all key metrics, including accuracy ($t(4) = 8.32, p = 0.0011$), macro F1 score ($t(4) = 7.15, p = 0.0020$), and AUC ($t(4) = 9.47, p = 0.0006$), the p-values are all well below the 0.01 significance level. This result strongly suggests that the performance advantages observed with the proposed method, compared to the current best

baseline model, are not due to random factors or specific data splits, but are highly statistically significant. Furthermore, we computed the effect size of the accuracy improvement, and this large effect size further confirms that the improvements achieved by the proposed method are substantial and practically meaningful.

4. CONCLUSION

This study addressed key issues in automatic diagnosis of respiratory system diseases and proposes an innovative framework based on multi-scale hierarchical feature fusion Transformer. By systematically combining multi-scale feature extraction, time-aware modeling, and bidirectional feature fusion mechanisms, the framework effectively solved the multi-scale representation problem of pathological features in respiratory audio signals. Experiments on multiple authoritative datasets including ICBHI, RespiratorySoundDB, and COVID-19 cough sounds demonstrated that the framework achieved breakthrough performance in respiratory sound classification, significantly outperforming traditional machine learning methods and mainstream deep learning models. Notably, the model showed excellent data efficiency, maintaining high relative performance with only 50% training data, demonstrating strong generalization ability and clinical applicability. Ablation experiments further validated the effectiveness of the multi-scale feature fusion strategy, proving its ability to collaboratively utilize comprehensive information from millisecond-level audio events to minute-level breathing patterns.

The main contribution of this study is the first systematic introduction of the multi-scale hierarchical fusion concept into the field of respiratory audio analysis, providing a new technical pathway for intelligent diagnosis of respiratory system diseases. The proposed framework not only achieves excellent diagnostic performance but also offers good interpretability, as model decisions can be visualized through attention maps, enhancing clinical trustworthiness. However, the study has some limitations: first, model training relies on high-quality annotated data, which is costly in the medical field; second, the framework has relatively high computational complexity, posing challenges for deployment in resource-limited environments; third, the current study mainly focuses on common respiratory system diseases, and the diagnostic ability for rare diseases requires further verification. Future research will focus on the following directions: First, we will explore contrastive learning-based self-supervised pretraining strategies, using a large amount of unlabeled respiratory audio data to construct pretraining tasks, aiming to enhance the model's generalization ability in low-sample scenarios. Second, we will study model lightweighting and embedded deployment solutions to enable compatibility with mobile devices and edge computing nodes, facilitating home-based screening and real-time monitoring of respiratory diseases. Additionally, we will expand the model's ability to fuse multimodal data, such as combining pulmonary function parameters and clinical questionnaire information, to build a more comprehensive respiratory health assessment system.

REFERENCES

[1] Redlarski, G., Jaworski, J. (2013). A new approach to

- modeling of selected human respiratory system diseases, directed to computer simulations. *Computers in Biology and Medicine*, 43(10): 1606-1613. <https://doi.org/10.1016/j.compbmed.2013.07.003>
- [2] Erdoğan, R.U., Kılıç, T., Çolak, T.K. (2024). Respiratory tract diseases with musculoskeletal system interaction: A scoping review. *Clinical and Experimental Health Sciences*, 14(2): 469-475. <https://doi.org/10.33808/clinexphhealthsci.1364053>
- [3] Oulefki, A., Agaian, S., Trongtirakul, T., Benbelkacem, S., Aouam, D., Zenati-Henda, N., Abdelli, M.L. (2022). Virtual Reality visualization for computerized COVID-19 lesion segmentation and interpretation. *Biomedical Signal Processing and Control*, 73: 103371. <https://doi.org/10.1016/j.bspc.2021.103371>
- [4] Eren, Z.B., Vatansever, C., Kabadayı, B., Haykar, B., et al. (2024). Surveillance of respiratory viruses by aerosol screening in indoor air as an early warning system for epidemics. *Environmental Microbiology Reports*, 16(4): e13303. <https://doi.org/10.1111/1758-2229.13303>
- [5] Karaarslan, O., Belcastro, K.D., Ergen, O. (2024). Respiratory sound-base disease classification and characterization with deep/machine learning techniques. *Biomedical Signal Processing and Control*, 87: 105570. <https://doi.org/10.1016/j.bspc.2023.105570>
- [6] Yu, G., Yu, Z., Shi, Y., Wang, Y., et al. (2021). Identification of pediatric respiratory diseases using a fine-grained diagnosis system. *Journal of Biomedical Informatics*, 117: 103754. <https://doi.org/10.1016/j.jbi.2021.103754>
- [7] Agarkov, S.F. (1999). Deficiency of conditioning function of the respiratory system in some respiratory and circulatory diseases. *Terapevticheskii Arkhiv*, 71(3): 48-51.
- [8] Han, L., Liang, W., Xie, Q., Zhao, J., Dong, Y., Wang, X., Lin, L. (2023). Health monitoring via heart, breath, and korotkoff sounds by wearable piezoelectret patches. *Advanced Science*, 10(28): 2301180. <https://doi.org/10.1002/advs.202301180>
- [9] Rocha, B.M., Filos, D., Mendes, L., Serbes, G., et al. (2019). An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement*, 40(3): 035001. <https://doi.org/10.1088/1361-6579/ab03ea>
- [10] Fan, D., Yang, X., Zhao, N., Guan, L., et al. (2024). A contactless breathing pattern recognition system using deep learning and WiFi signal. *IEEE Internet of Things Journal*, 11(13): 23820-23834. <https://doi.org/10.1109/IJOT.2024.3386645>
- [11] Urtnasan, E., Park, J.U., Lee, K.J. (2020). Automatic detection of sleep-disordered breathing events using recurrent neural networks from an electrocardiogram signal. *Neural computing and applications*, 32(9): 4733-4742. <https://doi.org/10.1007/s00521-018-3833-2>
- [12] Shayegh, S.V., Tadj, C. (2025). Deep audio features and self-supervised learning for early diagnosis of neonatal diseases: Sepsis and respiratory distress syndrome classification from infant cry signals. *Electronics*, 14(2): 248. <https://doi.org/10.3390/electronics14020248>
- [13] Monge-Álvarez, J., Hoyos-Barceló, C., San-José-Revuelta, L.M., Casaseca-de-la-Higuera, P. (2018). A machine hearing system for robust cough detection based on a high-level representation of band-specific audio features. *IEEE Transactions on Biomedical Engineering*,

- 66(8): 2319-2330.
<https://doi.org/10.1109/TBME.2018.2888998>
- [14] Rivas-Navarrete, J.A., Pérez-Espinosa, H., Padilla-Ortiz, A.L., Rodríguez-González, A.Y., García-Camero, D.C. (2025). Edge computing system for automatic detection of chronic respiratory diseases using audio analysis. *Journal of Medical Systems*, 49(1): 1-19. <https://doi.org/10.1007/s10916-025-02154-7>
- [15] Silva, L., Valadão, C., Lampier, L., Delisle-Rodríguez, D., Caldeira, E., Bastos-Filho, T., Krishnan, S. (2022). COVID-19 respiratory sound analysis and classification using audio textures. *Frontiers in signal processing*, 2: 986293. <https://doi.org/10.3389/frsip.2022.986293>
- [16] Mukherjee, H., Sreerama, P., Dhar, A., Obaidullah, S.M., Roy, K., Mahmud, M., Santosh, K.C. (2021). Automatic lung health screening using respiratory sounds. *Journal of Medical Systems*, 45(2): 19. <https://doi.org/10.1007/s10916-020-01681-9>
- [17] Ntalampiras, S. (2023). Explainable Siamese neural network for classifying pediatric respiratory sounds. *IEEE Journal of Biomedical and Health Informatics*, 27(10): 4728-4735. <https://doi.org/10.1109/JBHI.2023.3299341>
- [18] Pessoa, D., Rocha, B.M., Gomes, M., Rodrigues, G., et al. (2024). Ensemble deep learning model for dimensionless respiratory airflow estimation using respiratory sound. *Biomedical Signal Processing and Control*, 87: 105451. <https://doi.org/10.1016/j.bspc.2023.105451>
- [19] Mukherjee, H., Salam, H., Santosh, K.C. (2021). Lung health analysis: Adventitious respiratory sound classification using filterbank energies. *International Journal of Pattern Recognition and Artificial Intelligence*, 35(14): 2157008. <https://doi.org/10.1142/S0218001421570081>