Mel-TansWeD-CNN: Respiratory Diseases Detection Framework Through Acoustic Features and Fusion of CNN with Transformer

Auliya Ur Rahman, Asima Ismail
University of Engineering and Technology Taxila.
*Correspondence: auliyaurrahman555@gmail.com

Citation | Rahman. A. U, Ismail. A, "Mel-TansWeD-CNN: Respiratory Diseases Detection Framework Through Acoustic Features and Fusion of CNN with Transformer", IJIST, Vol. 8 Issue. 4 pp 1604-1619, July 2026
Received | June 08, 2026 **Revised |** July 07, 2026 **Accepted |** July 10, 2026 **Published |** July 15, 2026.

Respiratory diseases represent a significant global health challenge, which requires a reliable framework for early diagnosis detection framework to assist clinicians in making their clinical decisions. Elderly people and people living in industrial areas are commonly most affected due to pollution. In this study, we designed Mel-TransWed-CNN, a hybrid deep learning (DL) framework that fuses Mel frequency cepstral coefficients (MFCC) with Transformer and Convolutional Neural Network (CNN) to capture local spectral and global temporal patterns, respectively. We used the standard ICBHI 2017 respiratory sound dataset for experiments. The proposed approach consists of three main stages. First, the raw audio recordings are resampled and normalized to ensure consistency and improve signal quality. Next, MFCC with 20 coefficients are extracted to capture discriminative acoustic characteristics of respiratory sounds. Afterward, the extracted features are fed into the proposed fusion architecture, which combines a CNN with a transformer, where an initial CNN layer learns local spectral patterns, followed by a transformer encoder that models long-range temporal dependencies. A subsequent CNN stage further refines representations before pooling and Softmax-based multiclass classification. Furthermore, the proposed Mel-TransWeD-CNN achieved a mean accuracy of 95% with a standard deviation of 0.64% and a 95% confidence interval of 92.91%–95.43%, while outperforming the baseline classifiers, including Naive Bayes (88.0%), Gaussian Process Classifier (87.5%), Nearest Centroid (77.9%), and AdaBoost (71.7%). The proposed method contributes a practical and reliable automated framework for respiratory disease detection, with potential for early diagnosis and remote patient monitoring.

Keywords: Respiratory diseases, CNN, Transformer, Encoder, MFCC, Lung sounds

Introduction:

Respiratory diseases represent a major class of pulmonary diseases that affect air pathways and lung tissues, which contribute significantly to global mortality. Early diagnosis is effective for disease management as it reduces overall cost. The World Health Organization (WHO) identifies different respiratory diseases such as chronic obstructive pulmonary disease (COPD), asthma, acute lower respiratory tract infection (LRTI), tuberculosis, and lung cancer [1]. The number of affected people around the world by COPD is more than 65 million and more than 3 million deaths occur every year, which is the third leading cause of death around the world. Asthma is also a very common disease that affects approximately 339 million people around the world, and it is considered a chronic childhood disease. The common disease that affects children of ages lower than 5 years is pneumonia. Moreover, lung cancer also kills over 1.6 million people per year. There are other diseases of the lungs such as crackle, rale, rhonchi, and wheeze, which are very common around the world. These diseases are avoidable and curable. Early prevention and detection are very significant to limit the rapid spread of these life-threatening diseases.

DL architectures have increasingly been explored for automated respiratory disease diagnosis through acoustic signal analysis. Multi-path and multi-branch network architectures, though predominantly utilized in medical image segmentation, have recently gained attention in audio-based classification tasks, particularly for distinguishing complex respiratory pathologies from auscultation recordings. Medical experts are specially trained to carefully listen and recognize the pathologic findings such as the existence of disorders (like crackle, wheeze, or both, etc.,). Even though the auscultation devices have grown from analog to digital, therefore making enables to store, evaluate, and visualize in computer machines, however, digital auscultation isn't considered a mature technique. There are some limitations of traditional auscultation that restrict its use in the research due to the reasons such as continuous monitoring is not possible, experts are required for the detection of the disease. [2] proposed a Swin Transformer-based architecture to overcome CNN limitations in modeling long-range dependencies in respiratory sounds. Using MFCC inputs and hierarchical shifted-window blocks, the model captured both local and global acoustic features. It achieved 88.5% accuracy and an F1-score of 88.3%, but still, there exist challenges to discriminate acoustically overlapping respiratory conditions. [3] proposed a thermal imaging-based COPD classification system extracting four respiratory features. The features were scaled using Z-score normalization and combined through weighted summation with ROC curve-based thresholding. The approach achieved 85% accuracy. However, the limited sample size limits the generalizability of the approach to diverse patient populations. [4] proposed a four-layer CNN for respiratory disease classification using augmented electronic stethoscope recordings. Multiple data augmentation techniques were applied to enhance training diversity, which resulted in 90% accuracy. However, the study did not report class-wise performance metrics or clearly specify the number of disease categories, which has limitation of comprehensive evaluation. [5] developed a transfer learning-based CNN model for multi-class respiratory sound classification using 1,918 clinical recordings. The model achieved 86.5% accuracy and an AUC score of 0.93 for abnormal detection while for an accuracy of 85.7% for multi-class classification was obtained. However, it has also limitations because the performance degradation was observed in mixed abnormal sounds and noisy clinical environments.

[6] employed EfficientNetV2-M with transfer learning for multi-class lung disease classification from chest X-ray images. The model achieved approximately 82% accuracy across two different datasets. However, significant inter-class variability and sensitivity to dropout regularization affected consistent performance of the approach. [7] introduced a blockchain-integrated federated learning framework for multi-class respiratory disease classification. The proposed aggregation strategy achieved 88.1% accuracy for five-class

classification. Although the model achieved an accuracy of 88.1%, but the performance of the model depended heavily on heterogeneous local data quality and probable overfitting. [8] implemented a hybrid model consisting of long short-term memory (CNN-LSTM) architecture with focal loss using STFT spectrograms from the ICBHI 2017 dataset. The model achieved up to 76.39% accuracy using cross-validation technique. However, sensitivity remained relatively low, which indicates challenges in balanced abnormality detection. [9], two features, namely, STFT and wavelet-based feature fusion was combined and an SVM classifier was utilized for four-class respiratory sound classification, which achieved 49.86% accuracy. However, the exist limitations of the conventional feature engineering approaches. [10], a noise-masking recurrent neural network (NMRNN) was proposed for multi-class anomaly detection in lung sounds. The model achieved 56% sensitivity and 73.6% specificity, but detection capability remained moderate compared to advanced deep architectures, so, this approach was also not reliable to be employed for purpose of respiratory disease detection. In [11], a CNN-based model was developed for crackle and wheeze detection using preprocessed lung sounds. The approach achieved 50% accuracy under different train-test splits, but the results clearly indicate very limited generalization capability. A hybrid CNN-RNN framework utilizing MFCC inputs was proposed in [12] to capture spatial and temporal features. The CNN-RNN approach achieved an accuracy of 66.31%. However, performance remained degraded for reliable clinical deployment. [13] employed a CNN with two features, namely, MFCC and power spectral density for crackle and wheeze detection. The best performance reached 43% accuracy under fivefold cross-validation. However, the outcomes of this approach also confirm significant limitations in discriminative capability and unreliability for detection of respiratory sounds.

Authors have applied Transformer-based and hybrid CNN Transformer architectures to respiratory sound analysis. [14] proposed an event-based transformer for respiratory phase and adventitious sound detection on the HF_Lung_V1 dataset, which achieves F1-scores of 78.9% for continuous adventitious sounds and 59.4% for discontinuous adventitious sounds. However, the approach is designed for sound-event detection rather than disease-level classification, so it limits its direct applicability to multi-class diagnosis. Similarly, [15] developed a multi-stage hybrid CNN-Transformer framework for pediatric lung sound classification using scalogram representations and class-wise focal loss, which secured an overall score of 84.48% for multiclass event classification and only 57.1% at the recording level. However, the model was evaluated specifically on a pediatric population, which limits generalizability to adult and elderly patients who represent the majority of respiratory disease cases. In another study, [16] introduced a CNN with temporal self-attention combined with a frequency band selection module, it resulted in an accuracy score of 83.49% on the SPRSound 2023 dataset. However, performance remained dataset-dependent, with results varying across pediatric and adult respiratory sound corpora. [17] proposed a Multi-View Spectrogram Transformer that embeds multiple time-frequency views of the mel-spectrogram into a vision transformer by means of a gated fusion scheme, improved performance over prior state-of-the-art methods on the ICBHI 2017 dataset. However, the approach relies on pretrained AudioSet weights for initialization, which may limit reproducibility in settings without access to large-scale pretraining resources.

Besides Transformer-based approaches, other recent studies have explored recurrent and clustering-based architectures for respiratory sound classification. [18] proposed a lightweight framework based on improved deep clustering and contrastive learning, secure an ICBHI score of 63.26% with a sensitivity of 44.47% and specificity of 82.06%. However, the relatively low sensitivity indicates continued difficulty in correctly identifying abnormal respiratory sound classes. Furthermore, authors have explored the class imbalance problem inherent in the ICBHI 2017 dataset, which is directly relevant given the random oversampling

(ROS) strategy adopted in this work. [19] employed Variational Autoencoders, which include multilayer Perceptron VAE and Convolutional VAE variants, to synthesize minority-class respiratory sounds, resulting in improved classification performance for underrepresented classes following augmentation. However, the quality of synthetic samples was found to vary across classes, with only marginal improvement observed for some minority categories. [20] planned a deep CNN combined with artificial noise addition for abnormal respiratory sound classification. However, the approach was evaluated primarily on binary abnormal-versus-normal classification rather than a full multi-class disease taxonomy. [21] presented an input-agnostic representation-level augmentation technique for respiratory sound classification. However, the method requires careful tuning of augmentation strength to avoid degrading minority-class representations. [22] developed a lightweight inception network for respiratory disease classification using lung sounds. However, the reduced model complexity involved a tradeoff with discriminative capacity for acoustically similar disease classes. Similarly, [23] proposed patch-mix contrastive learning combined with an audio spectrogram transformer for respiratory sound classification. However, the method's reliance on large-scale incompatible pretraining increases computational cost compared to conventional supervised training. Furthermore, [24] proposed a dual-channel CNN-LSTM algorithm for lung sound classification using data augmentation and sampling techniques, which results in high accuracy on an augmented dataset. However, the significant performance gain was closely linked to the specific augmentation and sampling pipeline used, raising questions about generalization to differently distributed datasets.

Other than classification accuracy, recent work has also emphasized the computational feasibility and clinical deploy ability of respiratory sound classification systems. [25] explored DL-based audio enhancement to improve the robustness and clinical applicability of automatic respiratory sound classification, which results in inference times of 1.5 to 26 milliseconds per second of audio depending on the enhancement architecture used. However, most audio enhancement techniques were originally optimized for speech rather than respiratory sounds. [26] reviewed recent progress in lightweight model design and hardware platform implementation for intelligent lung sound diagnosis, which highlights solutions such as MobileNetV2-based classifiers that achieve real-time inference on embedded and mobile devices. However, the review notes that lightweight architectures often involve a tradeoff between reduced computational cost and discriminative capacity for closely related disease classes.

The lung sounds have a non-linear and non-stationary nature, which makes the detection significantly inflexible for unhealthy cases. Therefore, we need to design an effective respiratory disease detector that is capable of capturing precisely the non-linear and non-stationary nature of the lung sounds. Although the approaches discussed in the above literature review have employed different DL and machine learning (ML) approaches using spectral features for respiratory sound analysis, however, the existing ML and DL approaches show limitations in performance to detect the multi-class problem of respiratory sounds. To overcome these limitations, this proposed approach describes the following objectives and novel contributions, we propose a Mel-TransWed-CNN architecture that fuses MFCC features with Transformer and CNN-based classifiers for better respiratory disease classification. Respiratory sound analysis is critical for respiratory disease classification; however, the existing methods struggle with respiratory sound recognition due to its non-linear and non-stationary nature. To overcome these limitations, we proposed an advanced preprocessing and feature engineering pipeline that consists of noise handling and discriminative MFCC representation to improve acoustic pattern learning. Furthermore, we performed extensive experiments and compared our proposed model with conventional ML

and state-of-the-art DL methods to validate the reliability and effectiveness of the proposed model.

As discussed above, previous studies have used CNN, RNN, and Transformer-based methods for respiratory sound classification, the proposed Mel-TransWed-CNN offers a unique contribution, Like the existing hybrid models, which uses Transformer or RNN after convolutional feature extraction, the proposed framework fuses CNN and Transformer in a layered structure, where the initial CNN layers extract local spectral patterns, a Transformer encoder models global temporal-frequency dependencies, and a subsequent CNN stage refines the representation before classification; furthermore, the previous studies evaluate less number of classes, the proposed model classify eight class classification on the ICBHI 2017 dataset making the classification more challenging. Also, the proposed approach addresses class imbalance through ROS in the MFCC feature space before training. As a result, the proposed method achieves 95% accuracy, exceeding the previous best reported result of 88.5%. The rest of the paper is structured as follows. Section II introduces the proposed methodology, Section III reports the dataset description and experimental analysis, and the final section IV concludes our work.

Material and Methods:

The proposed framework performs multiclass respiratory disease classification using a hybrid CNN–Transformer architecture, as illustrated in Figure 1. In the first stage of this approach, respiratory audio signals are first converted into time–frequency representations, which are used as 2D inputs to the CNN for extracting local acoustic patterns. Next, the resulting feature maps are reshaped into sequential form and processed by a Transformer encoder to capture long-range temporal and spectral dependencies. The encoder output is then reshaped back into a 2D representation and refined through additional CNN layers to enhance high-level feature learning. Finally, Global Average Pooling generates a compact feature vector that is passed to a Soft-Max classifier for disease classification. This fusion strategy combines conventional feature extraction with self-attention mechanisms, which improves the model's ability to discriminate complex respiratory sound patterns.

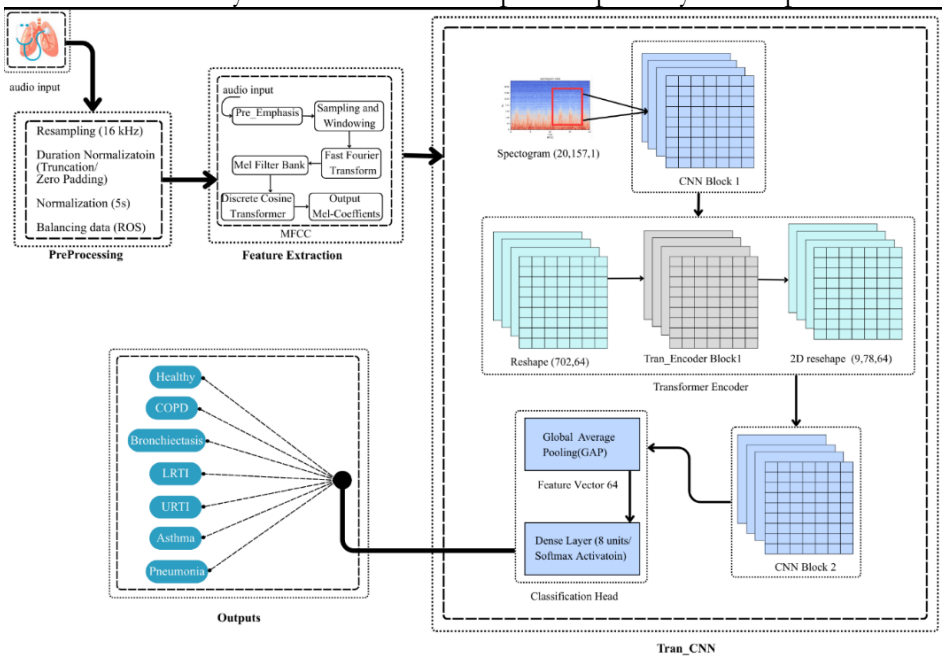

Figure 1. Proposed Working Mechanism.

Preprocessing

In this stage, we preprocess the raw audio samples such that we can get a uniform and noise-resistant representation as an input for the Mel-TransWed-CNN. Each input sample is

resampled to 16 kHz. The use of a 16 kHz sampling rate is well established in the speech processing literature, as it satisfies the Nyquist criterion for capturing the full human vocal range up to 8 kHz, and comparative studies have validated that this sampling rate paired with 16–24 MFCC coefficients achieves optimal recognition accuracy [27][28]. The signal was then normalized through zero-padding or truncation to a fixed duration of 5 seconds. This temporal window is widely adopted in speech emotion recognition tasks, as it preserves sufficient prosodic and contextual information for robust classification while maintaining computational efficiency [29]. The signal x(t) represents the raw audio sample or x(t) represents the raw audio sample, and the normalized signal is given by Eq. 1.

$$x'(t) = \begin{cases} [x(t), 0, \dots, 0] & \text{if } |x| < T \\ x(t)_{0:T} & \text{if } |x| > T \end{cases}, T = 5 \times 16{,}000 \qquad (1)$$

In Eq. 1, x(t) represents the original respiratory audio signal, while x'(t) denotes the processed audio signal after zero-padding and truncation, |x| is the length of the input recording, and T represents the fixed signal length matching 5 seconds (80,000 samples at a sampling frequency of 16 kHz). For feature extraction, Twenty MFCC coefficients are used, a formulation that has been applied in respiratory disease detection from lung sounds and has demonstrated effective classification performance [30]. ROS was applied to address class imbalance, where minority-class MFCC features were augmented in the feature space in order to achieve a normalized distribution before training. These steps were performed so we can get normalized, duration-matched, and feature-rich MFCC spectrograms, and as a result, we get a robust input representation for the proposed fusion model.

Feature Extraction Using MFCC:

In this stage of the approach, after input sample normalization, the MFCC consisting of 20 coefficients features were extracted to get the auditory relevant spectral characteristics of the lung sounds as shown in Figure 2.

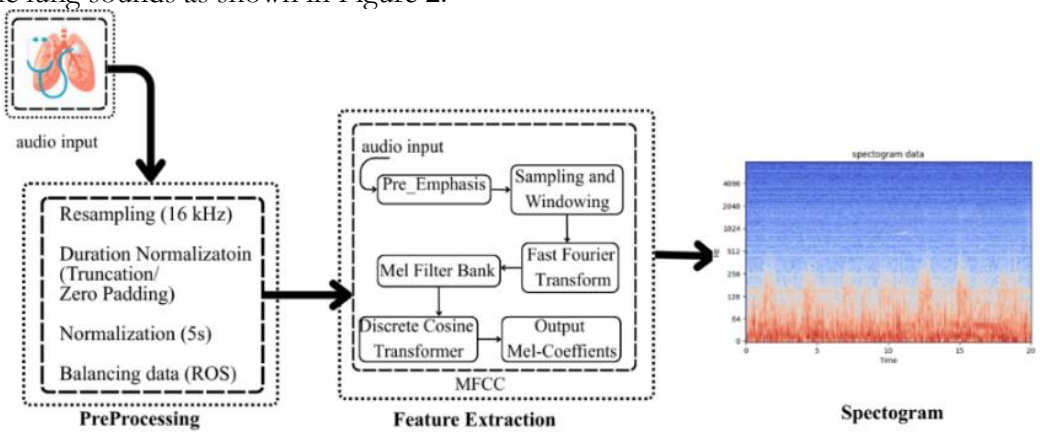

Figure 2. MFCC Spectrogram.

Short-Time Fourier Transform (STFT) is mapped onto the mel scale, and the discrete cosine transform is applied to it to get the MFCCs. This transformation yields a concise (20x157) time-frequency representation for each input sample, which is used as the primary input representation for the proposed Mel-TransWed-CNN. This representation is suitable for model processing because it captures the temporal, harmonic, and spectral dynamics, respectively, of lung sounds with reduced dimensionality.

$$\text{MFCC} = \text{DCT}\left(\log\left(\text{MelFilterBank}(|\text{STFT}(x'(t))|)\right)\right) \qquad (2)$$

In Eq. (2), (x'(t)) represents the preprocessed respiratory audio signal, STFT represents the Short-Time Fourier Transform, MelFilterBank applies Mel-scale filtering, log represents logarithmic scaling, and DCT denotes the Discrete Cosine Transform used to generate the MFCC features.

Classification:

In this stage, the proposed classifier integrates both convolutional learning and attention-based sequence modeling to perform multiclass respiratory disease prediction. The classification module is composed of two main sub-stages, namely, CNN-based feature learning and Transformer encoder–based contextual modeling, respectively. First, MFCC spectrograms are processed through initial convolutional layers to capture local spectral–temporal patterns and generate discriminative feature maps for further processing. These intermediate representations are then reshaped into token sequences and forwarded to a lightweight Transformer encoder, which models long-range dependencies using multi-head self-attention. After contextual refinement, the encoded features are reshaped back to spatial form and passed through additional CNN layers to enhance high-level feature abstraction. Global Average Pooling converts the final feature maps into a compact representation, which is fed into a fully connected SoftMax layer to produce class probabilities. The detailed structure of the Transformer Encoder Block and CNN Architecture used within this classification stage is described in the following subsections.

Transformer Encoder Block:

In the second stage of the model, we integrated a lightweight transformer encoder into the CNN layers, which allows it to capture long-range temporal-frequency dependencies that conventional CNN layers may often miss. After the first CNN layer, the feature map $X \in \mathbb{R}^{H\times W\times C}$ is reshaped into a sequence of (HW) tokens, each of dimension (C), using below Eq. 3.

$$X_{\text{seq}} = \text{Reshape}(H \cdot W, C) \qquad (3)$$

Where, X is the input feature map, $H$and Wdenote its height and width, C is the number of channels, and X_{seq} is the reshaped token sequence for Transformer input. The encoder uses multi-head self-attention, where each head computes it.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (4)$$

Here, Q, K, and Vdenote the query, key, and value matrices, respectively, $d_k$is the key dimension, and softmax computes the attention weights. So, the proposed model can learn long-range temporal frequency dependencies. We used an intermediate feed-forward layer of size 256, along with residual connections and dropout regularization to avoid overfitting and get a reliable transformation. For the next CNN layers, the sequence is reshaped back to its original spatial dimensions using Eq. 5.

$$X_{\text{out}} = \text{Reshape}(H, W, C) \qquad (5)$$

In above Eq (5), X_{out} is the reconstructed feature map obtained after reshaping the Transformer output. This enables the subsequent CNN layers to capture spatial features, which are enriched with global contextual information learned by the transformer. The fusion model combines the local patterns and long-range temporal-frequency dependencies in order to strengthen the representation of lung sounds.

CNN Architecture:

To extract local patterns from MFCCs hierarchically, we use 2D convolutional layers where each Convolutional operation implies a set of learnable filters over the input feature map, mathematically it is defined as under using Eq. 6.

$$Y(i,j,k) = \sigma\left(\sum_{m,n,c} X(i+m,j+n,c)\, W(m,n,c,k) + b_k\right) \quad (6)$$

Where (X) represents the input tensor, (W) shows the convolution kernels, (b_k) is the bias term, and sigma denotes the ReLU activation function. Furthermore, we down sample the feature maps using Eq. 7.

$$Y(i,j,k) = \max_{(m,n)\in\Omega} X(i+m, j+n, k) \quad (7)$$

In this equation, X and Y denote the input and output feature maps, respectively, (i, j) represents the spatial location, k is the feature channel, and Ω is the pooling window. The max-pooling operation performs down sampling by selecting the maximum activation within each pooling window. This down sampling makes the model robust to positional displacement in both time and frequency, thereby reducing computational complexity. To avoid overfitting, we applied dropout regularization after each convolution-pooling block by randomly deactivating a fraction of neurons during training. The final convolutional layer is pooled using Global Average Pooling (GAP) using Eq. 8.

$$z_k = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} X(i,j,k) \quad (8)$$

Here, z_k represents the pooled feature of the k-th channel, $X(i, j, k)$ denotes the feature value at spatial location (i, j) in channel k, and H and W represent the height and width of the feature map, respectively. The GAP compresses each feature map into a single descriptive value while preserving the channel-wise interpretability. Finally, the resultant vector is fed into a fully connected SoftMax layer for multiclass classification of lung sounds.

Fusion of Transformer and CNN:

The proposed fusion model unifies the benefits of convolutional processing and the self-attention mechanism through the transformer encoder block. Once the initial CNN layer captures the low-level spectral features, the intermediate feature map $X \in R\hat{}(H \times W \times C)$ is reshaped back into a sequence of tokens $X\,seq \in R\hat{}((HW) \times C)$, where each token preserves channel-wise and spatial information, which is used for improving global context reasoning. For the long-range dependencies, the transformer encoder uses multi-head self-attention, mathematically expressed as follows using Eq. 9.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (9)$$

In Eq. (9), Q, K, and V represent the query, key, and value matrices, respectively, and d_k denotes the dimension of the key vectors. The output is passed to a feed-forward network using Eq. 10.

$$\text{FFN}(x) = \sigma(xW_1 + b_1)W_2 + b_2 \quad (10)$$

Where, x is the input feature vector, W_1 and W_2 are learnable weight matrices, b_1 and b_2 are bias vectors, and σ denotes the activation function. The local details are captured first using CNN configuration and then the encoder captures global contextual information. Therefore, the robustness is improved from local to global long-range relational understanding. After learning acoustic features through Mel-TransWed-CNN, the final representation is passed into a fully connected classification layer. The Global Average Pooling converts the feature map into a concise low-dimensional feature vector $z \in R\hat{}C$. It is passed into a dense SoftMax layer to perform multiclass respiratory disease classification. The SoftMax layer calculates the probability of each class using Eq. 11.

$$P(y = c \mid z) = exp(z\_c) / sum\_\{k = 1\}\hat{}\{K\}\, exp(z\_k), \quad c = 1,2,\ldots,K \quad (11)$$

In Eq. (11), $P(y = c \mid z)$ denotes the probability of assigning the input to class c, z_c is the logit corresponding to class c, and K is the total number of respiratory disease classes. Categorical cross-entropy loss is used during training to adjust the network parameters and to improve the model ability to distinguish among different respiratory disease detections. This classification strategy ensures stable convergence and robust decision-making for complex lung sound patterns.

Algorithmic workflow:

For better clarity and reproducibility of the proposed framework, Algorithm 1 summarizes the complete workflow, from respiratory sound preprocessing to disease classification and performance evaluation.

Algorithm 1: Workflow of the Proposed Mel-TransWeD-CNN Framework.
Input: ICBHI 2017 respiratory sound recordings Output: Predicted respiratory disease class 1: Load the respiratory sound dataset. 2: for each audio recording do 3: Resample the signal to 16 kHz. 4: Normalize the signal to a fixed duration of 5 seconds. 5: Extract 20-dimensional MFCC features. 6: end for 7: Apply ROS to balance the training set. 8: Split the dataset into training and testing sets. 9: Train the Mel-TransWeD-CNN model: 10: Extract local features using CNN layers. 11: Learn global contextual representations using the Transformer Encoder. 12: Fuse CNN and Transformer features. 13: Classify respiratory diseases using the Fully Connected and Softmax layers. 14: Evaluate the trained model using Accuracy, Precision, Recall, F1-score, and Confusion Matrix. 15: Return the predicted respiratory disease class.

Result and Discussion:

The accurate and early diagnosis of respiratory diseases is a critical challenge in the field of medical diagnosis. In this work, we aim to design an approach named Mel-TransWed-CNN, which is built on acoustic features using the Mel-TransWed-CNN architecture, that uses lung sound for the detection of eight respiratory diseases, such as COPD, LRTI, URTI, and asthma, to overcome the limitations of existing work. We proposed a hybrid model Mel-TransWed-CNN, which consists of CNN layers as well as a transformer encoder for respiratory diseases classification, The first CNN layer captures the local acoustic patterns while the encoder captures the global temporal dependencies. Therefore, the fusion provides more discriminative classification of respiratory diseases. We explored various traditional ML approaches, such as Naive Bayes, Adaboost, Nearest Centroid, and Gaussian Process Classifier to compare our proposed model with these traditional approaches. The disease detector can detect multiple respiratory diseases at a time. Employing this proposed approach that can analyze complex lung sounds can be of great support for clinicians, which can be helpful in the enhanced and accurate detection of respiratory diseases.

Dataset:

In this work, we used the standard ICBHI dataset, namely, ICBHI, which consists of respiratory sound recordings from eight disease classes. The dataset contains recordings from the following classes, namely, asthma, bronchiectasis, bronchiolitis, COPD, healthy, URTI, LRTI, and pneumonia, respectively. Additionally, we trimmed each sound sample into 5-second segments to ensure all samples had a uniform duration, so, that the model can learn and capture the features. The dataset has two major subsets, namely, training and testing folders. The training folder contains training samples, which were used to train Mel-TransWed-CNN while the samples of the testing folder were used to evaluate the performance of the proposed approach. The traditional ML baselines (Naive Bayes, AdaBoost, Nearest Centroid, and Gaussian Process Classifier) were trained and evaluated using the identical training/testing split and the same 20-coefficient MFCC feature representation as the proposed Mel-TransWed-CNN model, ensuring a fair and consistent basis for comparison.

Experimental Setup:

The proposed Mel-TransWed-CNN framework was designed to extract a multi-level representation from the raw audio samples. Initially, the CNN initial layer captures local acoustic patterns, which is followed by the transformer encoder that captures long-range temporal dependencies across the spectrogram. Next, the CNN and global average pooling layers fuse the learned features and feed them to a dense classification head. We trained the model over 300 epochs and used the Adam optimizer along with categorical cross-entropy loss. We also used early stopping and model checkpointing to keep the best-performing weights. The graphs for training and validation accuracy curves are shown in Figure 3 (a) and 3 (b), respectively.

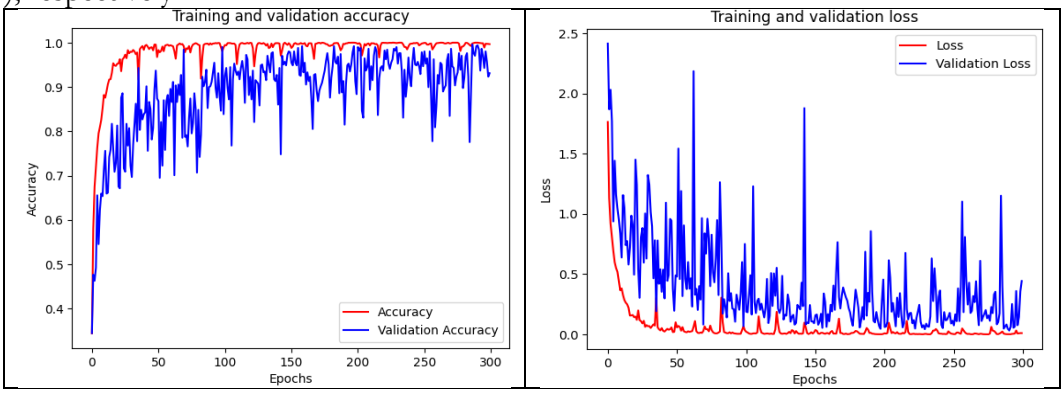

Figure 3. (a) Training and validation accuracy (b) Training and validation loss.

The curves represent smooth convergence in training metrics, while some fluctuations in validation loss show generalization across unseen data. Figure 3 (a) and 3 (b) illustrate the training dynamics of the proposed Mel-TransWed-CNN, over 300 epochs. Figure 3 (a) demonstrates rapid convergence, with training accuracy (red) reaching approximately 98% within 50 epochs and stabilizing near 100%, while validation accuracy (blue) exhibits characteristic fluctuations around 90–95%, indicating good generalization with only minor overfitting. Figure 3 (b) shows the corresponding loss curves, where training loss (red) decreases sharply and converges close to zero, while validation loss (blue) stabilizes at low values with periodic spikes, which reflect the model's learning progression and variance in validation data. The sustained gap between training and validation metrics suggests the model has learned robust feature representations while maintaining acceptable generalization capabilities for respiratory disease classification.

We evaluate the performance of Mel-TransWed-CNN using standard classification metrics, namely, accuracy, precision, recall, and F1 score, respectively. Additionally, we used confusion matrices to support class-wise evaluation and visualize true-positive and misclassification rates. We also used two other metrics, namely, area under the curve (AUC) and receiver operating characteristic (ROC) curves to analyze the discriminative ability of each class. Traditional ML models, namely, Naive Bayes (NB), Adaboost (AB), Nearest Centroid (NC), and Gaussian Process Classifier (GPC), respectively, were assessed based on weighted metrics to overcome class imbalance. The proposed classifier achieves excellent discriminative performance with an AUC of 1.00 for seven diseases, namely, Bronchiectasis, Bronchiolitis, COPD, Healthy, LRTI, Pneumonia, and URTI, respectively, while achieving an AUC of 0.99 for Asthma. The first confusion matrix (without class balancing) shows highly skewed predictions, with certain classes like COPD being overrepresented with 153 correct predictions while classes such as Asthma, Bronchiolitis, and Pneumonia, respectively, have very low true positive counts, which shows poor model generalization due to class imbalance as shown in Figure 4 (a).

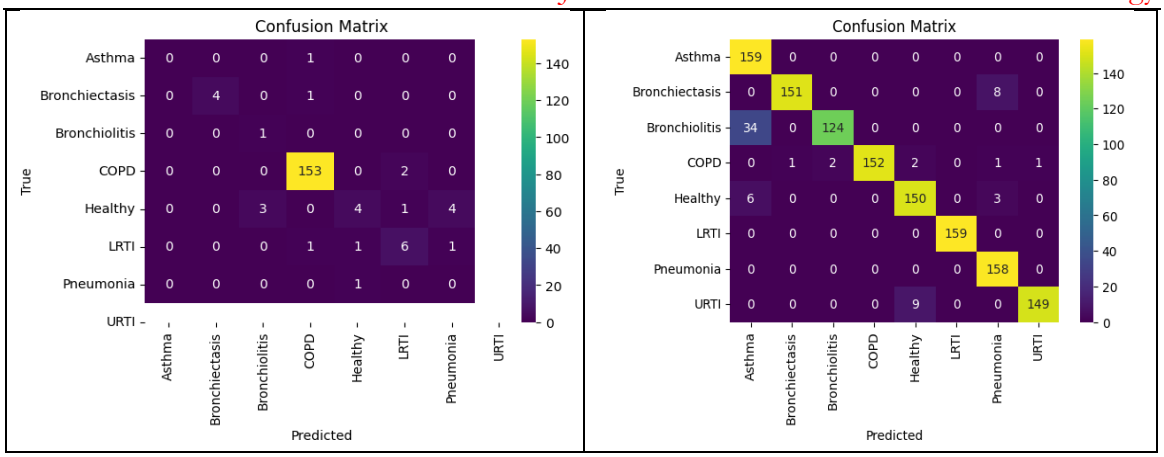

Figure 4: (a) CM Without Class Balance (b) CM After Balancing Classes.

Misclassifications are prominent for underrepresented classes, highlighting the model's bias toward dominant classes. The second confusion matrix, Figure 4(b), after class balancing, demonstrates significantly improved performance across all classes. True positives are much more evenly distributed, namely, Asthma has 159, Bronchiolitis has 124, and Pneumonia has 158, which show the model can now correctly classify minority classes. Misclassifications are reduced and mainly occur between similar respiratory conditions, which show that balancing the dataset moderates bias and enhances recall for rare classes, thereby improving overall classification reliability. The most prominent misclassification pattern occurs for Bronchiolitis, where 34 of 158 true Bronchiolitis samples were incorrectly classified as Asthma, while the remaining 124 were correctly identified, representing the lowest class-wise recall (78.5%) among all eight disease categories. This misclassification is acoustically reasonable, as both Bronchiolitis and Asthma are obstructive airway conditions that commonly present with wheeze-like adventitious sounds arising from narrowed or partially obstructed airways, which makes their spectral signatures more difficult to separate than those of structurally different conditions such as Pneumonia or LRTI. A smaller misclassification pattern is observed for Healthy samples, six of which were misclassified as Asthma. This reflects subclinical airway sounds present in some recordings labeled as healthy, which can share low-amplitude spectral characteristics with mild obstructive patterns. Similarly, three Healthy samples were misclassified as Pneumonia, and nine URTI samples were misclassified as Healthy, which suggests that upper-respiratory-tract sounds can, in some recordings, closely resemble normal breathing when infection severity or recording conditions reduce the distinction of adventitious sounds. In contrast, classes with more acoustically distinct signatures, namely Asthma, LRTI, and Pneumonia, achieved good recall, which represents that the proposed Mel-TransWed-CNN framework is most effective at separating conditions with clearly differentiated spectral-temporal patterns, while residual confusion remains concentrated among obstructive airway conditions with overlapping acoustic characteristics.

Comparison With Other Approaches:

In this experiment, we compared the performance of our approach, Mel-TransWed-CNN, with existing approaches. More specifically, Table 1 presents a comprehensive comparison of respiratory disease classification methods evaluated on the ICBHI 2017 dataset.

Table 1. Comparison of ICBHI 2017 Dataset Classification Results.

Method	Study	Classes	Accuracy (%)
Swin Transformer	[2]	6	88.50
CNN-LSTM + Focal Loss	[8]	4	76.39
SVM + STFT Wavelet	[9]	4	49.86
NMRNN	[10]	4	56.00*

CNN-RNN + MFCC	[12]	4	66.31

The proposed Mel-TransWed-CNN framework demonstrates superior performance and achieved an accuracy of 95% for eight-class respiratory disease classification. This represents a 6.5% improvement over the previous best reported result. [2] achieved 88.5% accuracy using Swin Transformer for six classes, while [8] obtained 76.39% accuracy with CNN-LSTM but showed a significant sensitivity-specificity imbalance of 52.78% vs. 82.46%, respectively. Traditional approaches showed considerably lower performance, for example, with [12] achieving 66.31% using CNN-RNN, and early CNN-based methods by [13] and [9] obtaining only 43% and 49.86%, respectively. The superior performance of Mel-TransWed-CNN demonstrates the effectiveness of combining CNN with a Transformer encoder. Moreover, local spectral pattern extraction with long-range temporal dependency modeling enables robust classification across eight respiratory disease classes with balanced precision-recall performance. Mel-TransWed-CNN outperformed all traditional ML models and achieved a classification accuracy of 95% on testing data.

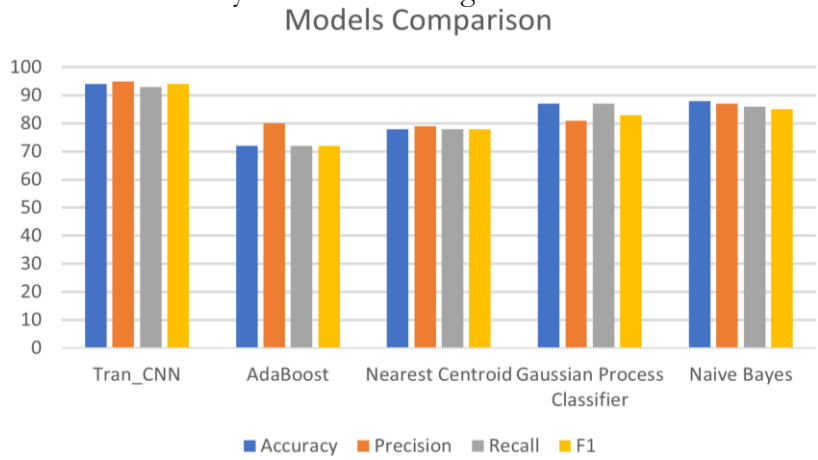

Figure 5. Model Comparison.

Figure 5 represents a performance comparison of Mel-TransWed-CNN with traditional classifiers using accuracy, precision, recall, and an F1-score. Mel-TransWed-CNN consistently achieves superior metrics, which shows robust classification capability. Furthermore, among the conventional methods, GPC and NB showed moderate performance, while AB and NC showed lower performance. Therefore, the comparative analysis of Mel-TransWed-CNN with traditional ML and existing approaches confirms that our approach is a better, more reliable, and more generalizable approach for automated respiratory disease classification, which can be employed in clinics to assist clinicians in their decision-making process.Next, we made comparative analysis of the results of Mel-TransWed-CNN with four traditional ML models, namely, NB, AB, NC, and GPC in terms of accuracy, macro- and weighted-average precision, recall, and F1-score, respectively as shown in Table 2.

Table 2. Comparison of Mel-TransWed-CNN and traditional ML Models.

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.88	0.89	0.88	0.88
AdaBoost	0.71	0.80	0.71	0.71
Nearest Centroid	0.77	0.78	0.77	0.77
Gaussian Process Classifier	0.87	0.81	0.87	0.83
Mel-TransWed-CNN	**0.95**	**0.95**	**0.95**	**0.95**

Mel-TransWed-CNN achieves the highest performance of 95% accuracy with macro-averaged precision, recall, and F1-score of 0.95 and weighted-average precision, recall, and F1-score of 0.95. However, traditional models achieve accuracies between 71%–88%. The results

confirm that classical models are unable to capture the complex temporal–spectral patterns in respiratory sounds, whereas our approach, Mel-TransWed-CNN, extracts local acoustic features and long-range dependencies. More specifically, the outcomes of our approach compared to other approaches showed that it has better feature learning and enhanced class separability, which reduces misclassification. The proposed Mel-TransWed-CNN achieved a mean test accuracy of 95% (standard deviation = 0.64%) with a 95% confidence interval of 92.91%-95.43%, obtained via bootstrap resampling (1,000 iterations) of the test set predictions. Notably, the lower bound of this confidence interval (92.91%) remains well above the accuracy achieved by the strongest traditional ML baseline, Naive Bayes (88.0%), indicating that the observed performance improvement of the proposed framework is unlikely to be attributable to sampling variability alone.

Ablation Study:

To evaluate the individual contribution of key components in the proposed framework, we performed a contribution analysis examining the effect of the ROS balancing strategy. Table 3 presents a comparison between the full Mel-TransWed-CNN model trained without class balancing and the complete proposed framework trained with ROS applied.

Table 3. Effect of class balancing (ROS).

Configuration	Accuracy	Macro-Avg. Recall
Without balancing	91.3%	54.10%
With balancing (proposed)	95%	94.31%

Removing the balancing approach caused only a modest reduction in overall accuracy (95% to 91.3%), which alone understates the true effect of class imbalance. Also, the macro-averaged recall, which weights all classes equally, dropped suddenly from 91.3% to 54.1%. This shows that the unbalanced model reaches apparently high accuracy mainly by correctly classifying the dominant COPD class (98.7% recall, n = 155), while minority classes such as Asthma and Pneumonia were almost entirely misclassified (0% recall for both). This approves that overall accuracy alone is an insufficient metric for imbalanced multi-class respiratory sound classification, and reveals that the ROS balancing strategy is essential not for borderline accuracy gain, but for reliable detection of underrepresented, clinically significant disease classes.

The strong performance of Mel-TransWed-CNN further reflects the balancing roles of its three core components. First, MFCC provides a compact, perceptually rich time-frequency representation that highlights the spectral signs most relevant to distinguishing pathological sounds such as crackles and wheezes from normal breathing, and has consistently outperformed raw waveform or basic spectrogram inputs in prior CNN- and Transformer-based classifiers [8][12]. Second, the initial CNN block extracts local, fine-grained spectral-temporal patterns corresponding to short-duration acoustic measures, consistent with previous work showing CNNs are well suited to such localized feature extraction [4][11][13], however, CNN-only architectures remain restricted in modeling dependencies across a full respiratory cycle. Third, the Transformer encoder directly addresses this limitation. Multi-head self-attention enables the model to relate acoustic measures occurring far apart in time, for example linking an early-cycle crackle with a later-cycle wheeze characteristic of a specific disease, which a local CNN receptive field cannot capture without excessive expansion. Recent hybrid CNN-Transformer approaches for respiratory sound classification have reported consistent gains over CNN-only baselines through this mechanism [2][14][15][17], reinforcing that global contextual modeling, rather than local feature extraction alone, is the primary contributor of improved class separability. By adding CNN and Transformer stages rather than applying attention only as a final step, as in prior single-pass hybrid designs [8][2], the proposed architecture allows locally extracted features to be globally contextualized and then spatially re-refined, which combines both mechanisms rather than treating them as

independent sequential stages. This complements the class-balancing contribution isolated empirically in Table 3, together validating each major design choice in the proposed framework.

Computational Complexity:

To evaluate the practical feasibility of deploying the proposed framework in real-time clinical situations, we evaluated its computational complexity and inference efficiency. The proposed Mel-TransWed-CNN contains a total of 68,120 trainable parameters, equivalent to a model size of approximately 266.09 KB, resulting in a model size smaller than large-scale DL architectures commonly used in medical image analysis. On the test set of 1,269 samples, the model achieved a total inference time of 0.69 seconds, which represents an average inference time of 0.55 milliseconds per sample. Given that each input sample represents a 5-second audio segment, this inference speed is several orders of magnitude faster than the input duration itself, indicating that the proposed framework is well suited for near real-time deployment on standard clinical computing hardware, like point-of-care devices with limited computational resources. The relatively compact model size and low inference latency, combined with the classification performance reported in Table 2, suggest that Mel-TransWed-CNN is a practical approach for use in remote patient monitoring systems and digital stethoscope applications without using significant computational or memory overhead.

Limitations:

Despite the good performance of the proposed Mel-TransWeD-CNN framework, several limitations persist. The model was evaluated on a single train/test split of the ICBHI 2017 dataset, which may limit its generalizability across different clinical settings and recording conditions. In addition, the proposed framework was validated only on eight respiratory disease classes and did not investigate the individual contribution of alternative acoustic feature representations. Future work will focus on validating the model on larger multi-center datasets and extending it into a multimodal framework by integrating respiratory audio with video and other clinical information to improve robustness and diagnostic performance

Conclusion:

This study presented, Mel-TransWed-CNN, a hybrid DL framework for multiclass respiratory disease classification using the ICBHI 2017 dataset. Respiratory diseases affect elderly people as well as younger people living in industrial areas. Therefore, we proposed a hybrid approach using 20-dimensional MFCC features for respiratory sounds representation and transformer integrated with CNN layers. More specifically, the proposed has capability to capture local acoustic characteristics and long-range temporal dependencies from MFCC-based respiratory sound representations. Experimental results illustrated that the proposed approach achieved superior performance compared to traditional ML models and existing approaches. More specifically, the Mel-TransWed-CNN has attained an overall accuracy of approximately 95% while it maintained strong class-wise evaluation metrics under class-imbalanced conditions. Therefore, our approach is reliable and generalizable for the timely and multiclass detection of respiratory disease, which can be employed in hospitals and clinics to make the decision-making process easier for clinicians. In future, we aim to explore the multimodal system using X-rays, respiratory videos, and respiratory sounds for more reliable detection of respiratory diseases.

Acknowledgement: The authors acknowledge the ICBHI 2017 dataset providers as its public availability.

Author's Contribution: All the authors have contributed equally.

Conflict of interest: Authors claim that there exists no conflict of interest for publishing this manuscript in IJIST.

References:

[1] "Automatic adventitious respiratory sound analysis: A systematic review - PubMed."

Accessed: Jul. 17, 2026. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28552969/

[2] W. Sun, F. Zhang, P. Sun, Q. Hu, J. Wang, and M. Zhang, "Respiratory Sound Classification Based on Swin Transformer," *2023 8th Int. Conf. Signal Image Process. ICSIP 2023*, pp. 511–515, 2023, doi: 10.1109/ICSIP57908.2023.10270823.

[3] "Classification of Chronic Obstructive Pulmonary Disease (COPD) Through Respiratory Pattern Analysis - PubMed." Accessed: Jul. 17, 2026. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/39941243/

[4] S. Sharma, S. Pandey, and D. Shah, "Enhancing Medical Diagnosis with AI: A Focus on Respiratory Disease Detection," *Indian J. Community Med.*, vol. 48, no. 5, pp. 709–714, 2023, doi: 10.4103/IJCM.IJCM_976_22.

[5] "Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning | Scientific Reports." Accessed: Jul. 17, 2026. [Online]. Available: https://www.nature.com/articles/s41598-021-96724-7

[6] "Deep Learning in Multi-Class Lung Diseases' Classification on Chest X-ray Images - PubMed." Accessed: Jul. 17, 2026. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35453963/

[7] A. Rahman *et al.*, "Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues," *Clust. Comput. 2022 264*, vol. 26, no. 4, pp. 2271–2311, Aug. 2022, doi: 10.1007/S10586-022-03658-4.

[8] Georgios Petmezas, Grigorios Aris Cheimariotis, "Automated Lung Sound Classification Using a Hybrid CNN-LSTM Network and Focal Loss Function," *Sensors*, vol. 22, no. 3, p. 1232, 2022, doi: https://doi.org/10.3390/s22031232.

[9] G. Serbes, S. Ulukaya, and Y. P. Kahya, "An automated lung sound preprocessing and classification system based onspectral analysis methods," *IFMBE Proc.*, vol. 66, pp. 45–49, 2018, doi: 10.1007/978-981-10-7419-6_8.

[10] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Noise masking recurrent neural network for respiratory sound classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11141 LNCS, pp. 208–217, 2018, doi: 10.1007/978-3-030-01424-7_21/SAVE-RESEARCH.

[11] "Lung Sounds (Breath Sounds): Types, Causes & Treatment." Accessed: Jul. 17, 2026. [Online]. Available: https://my.clevelandclinic.org/health/symptoms/25193-lung-sounds

[12] J. Acharya and A. Basu, "Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 3, pp. 535–544, Jun. 2020, doi: 10.1109/TBCAS.2020.2981172.

[13] P. Faustino, J. Oliveira, and M. Coimbra, "Crackle and wheeze detection in lung sound signals using convolutional neural networks," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2021, pp. 345–348, 2021, doi: 10.1109/EMBC46164.2021.9630391.

[14] J. Wang, G. Dong, Y. Shen, X. Xu, M. Zhang, and P. Sun, "REDT: a specialized transformer model for the respiratory phase and adventitious sound detection," *Physiol. Meas.*, vol. 13, no. 2, Feb. 2025, doi: 10.1088/1361-6579/ADAF08.

[15] Samiul Based Shuvo, Taufiq Hasan, "A Multi-Stage Hybrid CNN-Transformer Network for Automated Pediatric Lung Sound Classification," *arXiv:2507.20408*, Jul. 2025, Accessed: Jul. 22, 2026. [Online]. Available: https://arxiv.org/pdf/2507.20408

[16] N. Fraihi, O. Karrakchou, S. Member, and M. Ghogho, "Improving Deep Learning-based Respiratory Sound Analysis with Frequency Selection and Attention Mechanism," *arXiv:2507.20052*, Jul. 2025, Accessed: Jul. 22, 2026. [Online]. Available: https://arxiv.org/pdf/2507.20052

[17] W. He, Y. Yan, J. Ren, R. Bai, and X. Jiang, "Multi-View Spectrogram Transformer for Respiratory Sound Classification," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 8626–8630, 2024, doi: 10.1109/ICASSP48485.2024.10445825.

[18] Y. Chu, Q. Wang, E. Zhou, L. Fu, Q. Liu, and G. Zheng, "CycleGuardian: a framework for automatic respiratory sound classification based on improved deep clustering and contrastive learning," *Complex Intell. Syst. 2025 114*, vol. 11, no. 4, pp. 200-, Mar. 2025, doi: 10.1007/S40747-025-01800-4.

[19] J. Saldanha, S. Chakraborty, S. Patil, K. Kotecha, S. Kumar, and A. Nayyar, "Data augmentation using Variational Autoencoders for improvement of respiratory disease classification," *PLoS One*, vol. 17, no. 8, p. e0266467, Aug. 2022, doi: 10.1371/JOURNAL.PONE.0266467.

[20] R. Zulfiqar, F. Majeed, R. Irfan, H. T. Rauf, E. Benkhelifa, and A. N. Belkacem, "Abnormal Respiratory Sounds Classification Using Deep CNN Through Artificial Noise Addition," *Front. Med.*, vol. 8, p. 714811, Nov. 2021, doi: 10.3389/FMED.2021.714811.

[21] J.-W. Kim, M. Toikkanen, S. Bae, M. Kim, and H.-Y. Jung, "RepAugment: Input-Agnostic Representation-Level Augmentation for Respiratory Sound Classification," May 2024, Accessed: Jul. 22, 2026. [Online]. Available: https://arxiv.org/pdf/2405.02996

[22] A. Roy and U. Satija, "RDLINet: A Novel Lightweight Inception Network for Respiratory Disease Classification Using Lung Sounds," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, doi: 10.1109/TIM.2023.3292953.

[23] S. Bae *et al.*, "Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2023-August, pp. 5436–5440, Dec. 2024, doi: 10.21437/Interspeech.2023-1426.

[24] Y. Zhang, Q. Huang, W. Sun, F. Chen, D. Lin, and F. Chen, "Research on lung sound classification model based on dual-channel CNN-LSTM algorithm," *Biomed. Signal Process. Control*, vol. 94, p. 106257, Aug. 2024, doi: 10.1016/J.BSPC.2024.106257.

[25] J.-T. Tzeng *et al.*, "Improving the Robustness and Clinical Applicability of Automatic Respiratory Sound Classification Using Deep Learning-Based Audio Enhancement: Algorithm Development and Validation.," *JMIR AI*, vol. 4, no. 1, p. e67239, Mar. 2025, doi: 10.2196/67239.

[26] X. Lu, J. Fang, W. Xiao, and J. Wu, "Research progress and future prospects in intelligent lung sound diagnosis: models, lightweight design, and hardware platform implementation," *Biomed. Tech. (Berl).*, vol. 70, no. 6, pp. 483–501, Dec. 2025, doi: 10.1515/BMT-2025-0197.

[27] X. Mu and C. H. Min, "MFCC as Features for Speaker Classification using Machine Learning," *2023 IEEE World AI IoT Congr. AIIoT 2023*, pp. 566–570, 2023, doi: 10.1109/AIIOT58121.2023.10174566.

[28] "Speech Recognition: a review of the different deep learning approaches | AI Summer." Accessed: Jul. 23, 2026. [Online]. Available: https://theaisummer.com/speech-recognition/

[29] B. Zhu and L. Cao, "Implementation of the MFCC-Streaming Bi-LSTM Based Speech Emotion Recognition Method," *Proc. - 2025 Int. Conf. Artif. Intell. Electr. Electron. Eng. AI3E 2025*, pp. 109–114, 2025, doi: 10.1109/AI3E69313.2025.00030.

[30] C. Constantinescu and R. BRAD, "Low-Cost COPD Screening from Respiratory Sounds Using MFCCs with k-NN, SVM and Decision Trees," *Int. J. Integr. Eng.*, vol. 18, no. 4, pp. 130–138, Jun. 2026, doi: 10.30880/ijie.2026.18.04.009.