

Efficient ResNet-based Deep Learning Model for Asthma and COPD Detection Using Short-Time Fourier Transform Spectrograms

G. M. Karthik*

Department of Information Security, School of Computer science and engineering (SCOPE), Vellore Institute of technology, Vellore, TN, 632014, India

E-mail: karthik.gm@vit.ac.in

ORCID iD: <https://orcid.org/0000-0002-8425-7828>

*Corresponding author

D. Lakshmi Padmaja

Department of Information Technology, Anurag University School of Engineering, Hyderabad, 500088, India

E-mail: lakshmipadmajait@anurag.edu.in

ORCID iD: <https://orcid.org/0000-0001-8262-6821>

Nageswara Rao Medikundu

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, 522302, India

E-mail: medikundu1979@gmail.com

ORCID iD: <https://orcid.org/0000-0001-5668-2109>

C. Sateesh Kumar Reddy

Department of Advanced Computer Science and Engineering, Vignan's Foundation for Science, Technology and Research, Guntur, 522213, India

E-mail: satishreddic@gmail.com

ORCID iD: <https://orcid.org/0000-0002-1573-9073>

Pichili Vijaya Bhaskar Reddy

Department of Life Science and Bioinformatics, Assam University Diphu Campus, Diphu, 782462, India

E-mail: vpichili@gmail.com

ORCID iD: <https://orcid.org/0000-0002-9405-301X>

Anoop V.

Department of Artificial Intelligence and Data Science, Jyothi Engineering College, Thrissur, 679531, India

E-mail: vanoop@gmail.com

ORCID iD: <https://orcid.org/0000-0003-1960-8190>

Received: January 23, 2026; Revised: February 17, 2026; Accepted: March 23, 2026; Published: October 8, 2026

Abstract: This study proposes a deep learning-based framework for the automated detection of asthma and chronic obstructive pulmonary disease (COPD) using respiratory sound analysis. Breath sound recordings are preprocessed through resampling, silence trimming, and Wiener filtering to enhance signal quality. Short-Time Fourier Transform (STFT) is employed to convert audio signals into spectrogram representations, which are further augmented using time masking, frequency masking, and time warping to improve model generalization. The proposed model utilizes EfficientNet with multi-scale feature fusion to capture both local and global patterns in respiratory sounds. Experimental results demonstrate that the proposed approach achieves superior performance, with an accuracy of 99.21%, sensitivity of 99.56%, and specificity of 100%, outperforming existing CNN, ResNet, and SVM-based methods. The findings indicate that the proposed method is a reliable and efficient tool for non-invasive respiratory disease diagnosis.

Index Terms: Chronic Obstructive Pulmonary Disease, Wiener Filtering, Silence Trimming, Spectrograms, EfficientNet.

1. Introduction

Respiratory infections range from mild, self-limiting ailments like the common cold, influenza, and pharyngitis to more serious conditions like lung cancer, severe acute respiratory syndromes, bacterial pneumonia, pulmonary embolism, tuberculosis, acute asthma, and COPD, and can cause a wide range of symptoms. The main symptoms of COPD, a chronic, progressive respiratory disease, include breathing problems, increased sputum production, persistent coughing, and limitation of airflow. When COPD is at its worst, it drastically lowers quality of life, makes it harder to work, and can even cause disability. Over 300 million individuals worldwide suffer from COPD, and by 2030, this figure is predicted to increase even more [1].

First, early symptoms of COPD match those of other respiratory disorders, leading to frequent misdiagnoses or delayed discovery. Second, a large number of identified individuals do not receive adequate therapy, which hastens the course of the disease and causes serious consequences. Finally, the problem is made worse in some areas by the unequal distribution of medical resources, which hinders prompt diagnosis and treatment. New opportunities for the early detection and categorization of COPD have been made possible by recent developments in machine learning and deep learning. Particularly when processing and analyzing pulmonary sound data, machine learning algorithms examine enormous datasets to find trends and provide predictions. These methods help doctors distinguish between patient subgroups and diagnose COPD more accurately. Machine learning removes the subjectivity frequently seen in human judgment by developing and testing models on large respiratory sound datasets [2].

Support Vector Machines (SVM), Neural Networks (NN), and Random Forests (RF) are popular machine learning techniques. However, the lack of COPD samples and class imbalance may cause overfitting and misclassification in typical machine learning techniques. Deep learning methods perform better when managing complicated medical data than traditional machine learning techniques. They improve classification accuracy and diagnostic precision, making them useful tools for doctors to use when making decisions about pulmonary disease prediction. Convolutional neural networks (CNN) and long short-term memory (LSTM) networks are two examples of deep learning frameworks that have become well-known as important techniques for COPD diagnosis. These algorithms may efficiently extract important patterns without depending on conventional handcrafted features by utilizing massive datasets of respiratory sound recordings and spectrogram images. [3].

Historically, expert-driven knowledge-based systems or Pulmonary Function Tests (PFT) were used to diagnose COPD. These methods required meticulous image processing to accurately identify the condition. However, using chest X-rays for clinical diagnosis in emergency situations—like a heart attack, asthma attack, or severe respiratory distress—takes a lot of time and money and can even be dangerous. Furthermore, large datasets of high-resolution X-ray pictures are necessary for AI-driven image identification systems, but these can be challenging to reliably acquire. For quick COPD identification, a more approachable and resource-efficient method is needed to overcome these obstacles. An automated system that uses respiratory sound analysis to diagnose COPD is one potential method [4]. For diseases like heart disease, asthma, and COPD, the internal noises made by organs like the heart and lungs have important diagnostic significance. Both patients and medical professionals can profit from the very effective, self-sustaining diagnostic tool made possible by automating the categorization of these sounds. Future iterations of this system may include microphones and smart devices, allowing for continuous real-time monitoring of breathing noises to anticipate possible instances of COPD. Drug development, illness prediction, and medical imaging have all benefited greatly from deep learning, a specific area of machine learning built on artificial neural networks. Its use in lung sound classification continues to expand, giving automated feature extraction and training capabilities that minimize human effort while enhancing classification speed and accuracy. Designing the best deep learning architectures, improving model performance, and making the most of extracted features to increase diagnostic reliability are still difficult tasks [5].

Recent advances in deep learning, particularly convolutional neural networks (CNNs), have enabled automated feature extraction from spectrogram representations. Despite these improvements, most existing approaches utilize single-scale feature representations, limiting their ability to capture both fine-grained spectral variations (e.g., wheezes, crackles) and broader temporal patterns. Additionally, respiratory sound datasets often suffer from noise and variability, necessitating robust preprocessing and augmentation strategies. To address these limitations, this study proposes a multi-scale feature fusion framework based on EfficientNet for respiratory sound classification. The method integrates advanced preprocessing, STFT-based spectrogram generation, and task-specific augmentation techniques to enhance model robustness. By extracting and fusing features from multiple hierarchical layers, the proposed approach captures both local and global acoustic patterns, leading to improved classification performance.

The contribution of this work lies in the design of a unified deep learning framework that integrates multi-scale feature extraction within the EfficientNet architecture for respiratory sound classification. Unlike conventional approaches that rely primarily on single-scale or final-layer representations, the proposed method systematically fuses intermediate hierarchical features to capture both local and global acoustic patterns. This is complemented by a structured preprocessing and augmentation pipeline tailored for respiratory sound variability, resulting in improved robustness and classification performance.

The remainder of the article is organized as follows: Section 2 gives the literature review, Section 3 defines the methods and materials, Section 4 details the results, and lastly, Section 5 provides the conclusion.

2. Literature Survey

The use of artificial intelligence in medicine has been the subject of several clinical trials in recent years. Hospitalization rates, disease-related mortality, overall patient satisfaction, the frequency of medications given to patients, and changes in pulmonary function and their impact on quality of life have all been the subject of these studies. Patients with COPD have received special attention within this section [24].

The effectiveness of this approach was evaluated on an open lung sound database across different benchmark settings, showing strong statistical correlations with existing methods. Prakash et al. introduced a structured framework for lung sound data preprocessing, using noise filtering techniques to extract relevant features. After this preprocessing step, the model identifies and extracts key characteristics from lung sounds, which are then used to train the system based on these features. These features include the constant-Q transform, short-time Fourier transform, and Mel-frequency cepstral coefficients. The framework classified diseases including COPD, healthy lungs, upper respiratory tract infections, bronchiolitis, and pneumonia with 95.48% accuracy, 89.0% sensitivity, and 98.0% specificity [6]. Chen et al. introduced a unique approach for recognizing wheeze, crackle, and normal respiratory sounds by merging the optimal S-transform (OST) with deep residual networks (ResNets). The suggested OST is first used to analyze the raw respiratory sounds, producing spectrograms that are subsequently rescaled for the ResNet model to use. By improving feature extraction and classification, the deep residual learning method successfully differentiates between various respiratory sound patterns. This approach offers a dependable solution for telemedicine and e-health applications in respiratory illness detection as OST enhances the unique properties of wheezes, crackles, and other respiratory sounds while ResNet learns discriminative qualities. Furthermore, comparisons show that this approach outperforms empirical mode decomposition-based artificial neural networks (ANN) by 4.63% and deep learning-based ensemble CNN models by 3.23% in classification performance [7]. Mohammed presented the Optimal Asthma Disease Detection using Hybrid Machine Learning (OADD-HML) approach, an improved asthma detection system using speech inputs. By examining changes in the signal pattern in relation to normal and abnormal breathing frequencies, this method makes use of the Improved Weed Optimization (IWO) algorithm for the identification and prediction of asthma. Additionally, the study uses a Deep Q Neural Network (DQNN) for classification after employing an Enhanced Hunting Search (EHS) method for feature extraction and selection. Crackle noises, regular speech, and asthma are all recognized using the DQNN model. Based on measures for accuracy, sensitivity, and specificity, experimental evaluations showed how well the approach classified various respiratory diseases. OADD-HML performs better than current state-of-the-art methods when comparing asthma and non-asthma categories [8]. For the categorization of pulmonary diseases, Neili used a deep residual network (ResNet)-based model with three distinct architectures (ResNet-50, ResNet-101, and ResNet-152). As inputs for the ResNet models, the study used gammatonegram representations, which transform lung sounds from one-dimensional data into two-dimensional pictures. Three pulmonary conditions were classified using the ICBHI database: pneumonia, COPD, and healthy lungs. According to the results, the classification accuracies of ResNet-50, ResNet-101, and ResNet-152 were 90.37%, 89.79%, and 67.57%, respectively. These results demonstrate how deep residual networks may be used to reliably diagnose lung conditions [9].

To guarantee prompt medical action, Raghuram and Radhika underlined the significance of correctly diagnosing lung cancer and COPD. Their study concentrated on employing chest X-ray images for accurate illness categorization since differentiating between these two severe chest disorders is essential for successful therapy. Recurrent neural networks (RNNs) are used in the suggested strategy to improve radiograph feature extraction and classification accuracy. The program efficiently detects changes in lung architecture, including the development of tiny lobes or thickened capillaries in the respiratory system, by examining chest X-ray information. This method performs better in identifying and distinguishing respiratory disorders than traditional lung imaging approaches [10].

Despite significant advancements in machine learning and deep learning for respiratory sound analysis, several challenges remain unresolved. Existing approaches often rely on single-scale feature extraction, which limits their ability to simultaneously capture fine-grained acoustic signatures (such as wheezes and crackles) and broader temporal patterns present in respiratory signals. Furthermore, many models are sensitive to noise and variability in recording conditions, as they lack robust preprocessing and augmentation mechanisms. Another key limitation is that conventional CNN-based architectures typically utilize only high-level feature representations, potentially overlooking important intermediate features that contain valuable diagnostic information. Additionally, class imbalance and limited dataset diversity can lead to overfitting and reduced generalization performance in real-world clinical scenarios. Therefore, the research gap lies in the development of a robust, noise-resilient deep learning framework that effectively integrates multi-scale feature representations while improving generalization across diverse respiratory sound conditions. To address this gap, this study proposes a multi-scale feature fusion approach based on EfficientNet, combined with advanced preprocessing and spectrogram augmentation techniques, to enhance classification accuracy and reliability.

3. Methods

The suggested technique seeks to construct a robust asthma diagnosis system utilizing deep learning by exploiting multi-scale feature fusion with EfficientNet. The process begins with audio acquisition, where breath sounds are captured in .wav format. Given the inherent noise in real-world recordings, preprocessing is conducted to boost signal quality. In order to minimize background noise while maintaining crucial breathing patterns, Wiener filtering is used for adaptive noise reduction. Furthermore, resampling standardizes the data for uniformity across all samples, and quiet trimming eliminates superfluous pauses. The feature extraction stage uses STFT to transform the 1D audio samples into 2D spectrogram representations after preprocessing. The generated spectrograms are normalized to improve feature consistency and subsequently converted into a dB scale to increase contrast. Since deep learning models perform well with image-based inputs, the spectrograms are processed as pictures, enabling convolutional networks to capture detailed breathing patterns. Techniques for spectrogram augmentation are used to further improve the resilience of the model. By creating artificial occlusions in the spectrogram, time masking and frequency masking mimic real-world aberrations and enhance generalization. In order to keep the model invariant to slight temporal shifts in breath sounds, time warping is utilized to introduce variations in time. The heart of the model employs EfficientNet, a state-of-the-art CNN noted for its efficiency and great performance. Multi-scale feature fusion is used rather than just the last few layers of EfficientNet. To ensure that both local and global respiratory patterns contribute to the classification decision, features are extracted from multiple hierarchical blocks of EfficientNet, capturing both fine-grained and high-level representations. These feature maps are then resized to a common dimension using either convolutional downsampling or bilinear upsampling. The fused feature maps are then fused using weighted summation or concatenation, and a bottleneck projection layer (1x1 convolution) is added to reduce the dimensionality of the fused features while maintaining their vital characteristics. An adaptive weighting technique can then further improve the retrieved features once this compact representation has been run through global average pooling (GAP) or attention pooling. The last step involves applying a dense classifier with dropout and fully connected layers, followed by either a sigmoid function for binary classification (asthma vs. normal) or a softmax activation for multi-class classification. A final probability score that indicates whether a particular breath sound is indicative of a healthy person or an asthmatic is produced by the model. This method successfully combines deep learning-based classification, multi-scale feature extraction, and preprocessing to provide an asthma detection system that is both extremely accurate and effective. The suggested method's block diagram is displayed in Fig.1.

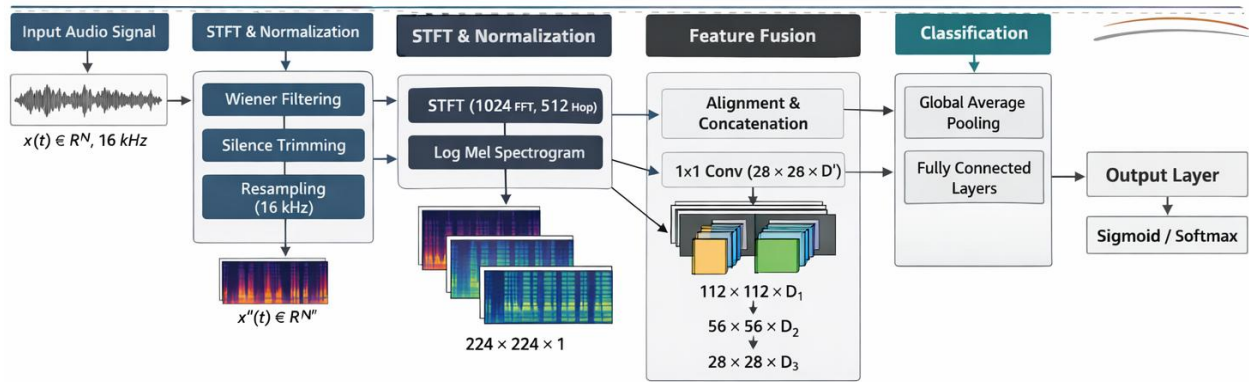


Fig. 1. Block diagram of the proposed method.

3.1. Preprocessing

Preprocessing plays a significant role in guaranteeing the quality and reliability of breath sound categorization. Feature extraction may suffer from the presence of background noise, silence, and fluctuating sample rates in the raw audio signals. Wiener filtering is used for adaptive noise reduction to overcome these problems, improving data clarity while maintaining essential breathing rhythms. Silence reduction reduces unwanted pauses, ensuring a more informative signal representation. After that, the audio is resampled (for example, to 16 kHz) to ensure uniformity over all samples. When taken as a whole, these actions increase the spectrogram features' quality and classification accuracy.

3.1.1. Wiener Filtering for Noise Reduction

A popular method for adaptive noise reduction, especially in speech and biomedical signal processing, is Wiener filtering [11,12]. The way it works is by reducing the mean square error (MSE) between the initial noisy signal and the anticipated clean signal. Let $X(f)$ be the observed noisy signal in the frequency domain. $S(f)$ be the clean signal. $N(f)$ be the noise component. $H(f)$ be the Wiener filter response.

$$H(f) = \frac{P_s(f)}{P_s(f) + P_N(f)} \quad (1)$$

where: $P_s(f)$ is the power spectral density (psd) of the clean signal, $P_N(f)$ is the PSD of noise. The estimated clean signal is then obtained by applying the filter:

$$\hat{S}(f) = H(f) \cdot X(f) \quad (2)$$

3.2. Feature Extraction

In this study, STFT is used to convert the 1D audio signal into a 2D spectrogram, capturing both time and frequency domain information. The spectrogram is then further normalized to a decibel (dB) scale to enhance feature contrast. The spectrogram is treated as an image to make the extracted features compatible with deep learning models, allowing CNNs to effectively learn spatial and spectral patterns. Feature extraction turns preprocessed breath sounds into meaningful representations for classification.

3.2.1. Short-Time Fourier Transform (STFT)

Breath sounds are examples of non-stationary signals that may be analyzed using the STFT, a basic approach that converts time-domain data to time-frequency domain signals [13,14,15]. The STFT divides the signal into small, overlapping segments called windows, allowing for a targeted frequency analysis over time, in contrast to the classic Fourier Transform, which only offers the frequency content of the full signal. In order to preserve time-dependent frequency fluctuations, each segment is then examined independently using the Fourier Transform. The STFT is mathematically defined as:

$$X(t, f) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j2\pi f\tau} d\tau \quad (3)$$

where: $X(t, f)$ represents the STFT output at time t and frequency f . $x(\tau)$ is the original time-domain signal. $w(\tau - t)$ is a window function applied to isolate a segment of the signal. $-j2\pi f\tau$ is the Fourier Transform operation applied to the windowed segment.

3.3. Normalization

In the context of lung sound analysis for asthma detection, normalization helps mitigate variations caused by different recording conditions, microphone sensitivities, and patient environments. Normalization is an essential preprocessing step in audio processing that ensures consistency in signal amplitude across different recordings [16,17,18]. It is especially pertinent when working with spectrogram-based machine learning and deep learning models, as inconsistent amplitude levels can negatively impact feature extraction and classification performance. One popular normalization technique is min-max normalization, in which the amplitude values are scaled to a fixed range, usually between 0 and 1 or -1 and 1 [19,20]. This keeps certain recordings from being overemphasized by using higher volume levels by guaranteeing that all audio samples have the same amplitude range. Normalization is frequently used for spectrogram-based representations in the decibel (dB) scale, where log compression is used to transform the spectrogram values to a logarithmic scale. This change lessens the prominence of louder components while highlighting lower-intensity noises, such as mild lung abnormalities.

3.4. Spectrogram Augmentation

In deep learning for audio and biomedical signal processing, spectrogram augmentation is a potent strategy that increases resilience to real-world fluctuations, decreases overfitting, and improves model generalization [13, 21]. It entails altering audio signal spectrogram representations to mimic various distortions, increasing the model's adaptability to a range of input circumstances. Applications such as voice recognition, music analysis, respiratory sound categorization, and asthma detection benefit greatly from this method. An audio signal's time-frequency representation created with STFT is called a spectrogram. Augmentation methods change these representations without changing the semantic content of the underlying audio signal since deep learning algorithms frequently regard spectrograms like pictures. The SpecAugment framework consists of the three most popular spectrogram augmentation techniques: time masking, frequency masking, and time warping.

Time masking is the process of choosing a section of the spectrogram at random along the horizontal axis of time and substituting it with a tiny constant or zero values. By simulating real-world distortions like signal dropouts or intermittent noise, this enables the model to learn to identify important audio patterns even when portions of the signal are absent. By masking the random frequency range, this strategy keeps the model from depending too much on any one frequency component. This is particularly helpful in biological audio applications, such as lung sound analysis, because various patient states naturally cause frequency changes. Time warping simulates changes in breathing patterns or speaking pace by softly compressing or stretching portions of the spectrogram along the time axis. This improves the model's capacity to accommodate varying respiration rates and speaking speeds.

3.5. Time Warping in Audio Processing

A popular data augmentation method in speech and sound analysis is time warping, which introduces temporal fluctuations in the input signal to improve model resilience [22]. Time warping alters the temporal dynamics of an audio signal without changing its frequency content, which is relevant for utilizing lung sound spectrograms for asthma identification. This helps ML and DL models generalize better by making them invariant to modest differences in breathing pace and rhythm. Time warping is the process of stretching or compressing specific spectrogram segments along the time axis by applying random time distortion or dynamic time warping (DTW). This guarantees that the model learns significant respiratory properties rather than overfitting to certain temporal trends [23]. Time steps within the spectrogram are often selectively deformed while preserving the general structure of the audio stream using spline-based interpolation or time interpolation methods. Time warping aids the model in learning invariant properties throughout the range of breathing rates, pauses, and abnormalities that patients may display while detecting asthma. For instance, the model will identify the same wheeze pattern even in the presence of temporal distortions if a spectrogram representation of a wheezing sound is significantly time-warped.

3.6. Mathematical Model for Multi-Scale Feature Fusion with EfficientNet

In the proposed framework, feature fusion is performed through concatenation of multi-scale feature maps, followed by dimensionality reduction using a 1×1 convolution. An attention mechanism is then applied to refine the fused representation by adaptively weighting spatial features. Weighted summation is not used as an alternative fusion strategy but as part of the attention-based refinement stage. Let the input spectrogram be denoted as:

$$X \in \mathbb{R}^{T \times F \times C} \quad (4)$$

Where T is the number of time steps, F number of frequency bins, C is the number of channels (1 for grayscale, 3 for RGB spectrogram). Let EfficientNet consist of L stages. We extract intermediate features from selected stages $l_1 < l_2 < l_3 \in \{1, 2, \dots, L\}$. The feature maps from each selected stage are:

$$F_{l_i} = E_{l_i}, i \in \{1, 2, 3\} \quad (5)$$

Where E_{l_i} is the transformation function (convolution, activation, pooling) up to stage l_i , $F_{l_i} \in \mathbb{R}^{T \times F \times D_i}$ is the extracted multi-scale features, D_i is the channel depth at stage l_i . To unify spatial resolution across scales, all feature maps are resized to a common spatial size (T', F') :

$$\tilde{F}_{l_i} = \text{Resize}(F_{l_i}) \in \mathbb{R}^{T' \times F' \times D_i} \quad (6)$$

$$F_{\text{fusion}} = \text{Concat}(\tilde{F}_{l_1}, \tilde{F}_{l_2}, \tilde{F}_{l_3}) \in \mathbb{R}^{T' \times F' \times (D_1 + D_2 + D_3)} \quad (7)$$

$$F_{\text{fusion}} = \sum_{i=1}^3 \alpha_i \cdot \tilde{F}_{l_i} \quad (8)$$

where $\alpha_i \in [0, 1]$, $\sum_{i=1}^3 \alpha_i = 1$. To compress the fused representation using a 1×1 convolution:

$$F_{\text{proj}} = \sigma(W_p * F_{\text{fusion}} + b_p) \quad (9)$$

where $W_p \in \mathbb{R}^{1 \times 1 \times (D_1 + D_2 + D_3) \times D'}$. D' is the reduced channel dimension, σ is the activation function (e.g., ReLU, Swish).

$$z = \frac{1}{T' \times F'} \sum_{t=1}^{T'} \sum_{f=1}^{F'} F_{\text{proj}}(t, f) \quad (10)$$

$$z = \sum_{t=1}^{T'} \sum_{f=1}^{F'} a_{t,f} \cdot F_{\text{proj}}(t, f) \quad (11)$$

$$a_{t,f} = \text{Softmax}(g(F_{\text{proj}}(t, f))) \quad (12)$$

where $g(\cdot)$ is a lightweight attention scoring network.

$$h = \phi(W_1 z + b_1) \quad (13)$$

$$y = \text{Softmax}(W_2 h + b_1) \quad (14)$$

$$y = \sigma(w^T h + b) \quad (15)$$

where $\phi(\cdot)$ is the non-linear activation (e.g., ReLU), $\sigma(\cdot)$ is the sigmoid function. The proposed model was implemented using Python with deep learning frameworks including PyTorch and Keras. The EfficientNet backbone was initialized with pretrained weights and fine-tuned on the respiratory sound dataset. Spectrograms were resized to 224×224 and normalized before being fed into the network. The model was trained using the Adam optimizer with an initial learning rate of 1×10^{-4} . A step decay scheduler was applied, reducing the learning rate by a factor of 0.1 after every 10 epochs without improvement. For binary classification (asthma vs normal), Binary Cross-Entropy loss was used. For multi-class classification, Categorical Cross-Entropy loss was applied. The model was trained with a batch size of 32 for 50 epochs. Early stopping with a patience of 8 epochs was employed to prevent overfitting. Dropout with a rate of 0.5 was applied in the fully connected layers. Data augmentation (time masking, frequency masking, and time warping) was also used to improve generalization. Experiments were conducted using multiple train-test splits (50:50 to 80:20), with the best performance observed at 80:20.

4. Results and Discussions

In this section, we report the experimental findings and performance analysis of our proposed Multi-Scale Feature Fusion EfficientNet model for asthma diagnosis utilizing spectrogram-based representations of breath sounds. The model was assessed using publicly accessible datasets, and its effectiveness was compared with existing deep learning architectures. A thorough performance review is ensured by the evaluation measures, which include accuracy, precision, sensitivity, F1-score, and specificity. Using cutting-edge deep learning techniques, the implementation employs a systematic workflow of preprocessing, feature extraction, augmentation, and classification. To ensure faster computing, the model was trained and evaluated using a Google Colab environment equipped with NVIDIA T4 GPUs. Python 3.7, PyTorch 2.10.0, and Keras 2.9.1 are included in the software environment, together with necessary libraries for data preparation and performance assessment such as NumPy, SciPy, Matplotlib, and Scikit-learn.

4.1. Asthma Detection Dataset: Version 2

The dataset used in this study consists of 1,211 respiratory sound recordings categorized into five classes: asthma (288 samples), bronchitis (104 samples), COPD (401 samples), pneumonia (255 samples), and healthy (133 samples). Each recording was segmented into durations ranging from 1.5 to 5 seconds to ensure consistency during model training. The dataset was curated from publicly available respiratory sound repositories and standardized to ensure uniform sampling conditions. All audio recordings were converted to a common sampling rate of 16 kHz and stored in .wav format. Preprocessing steps, including noise filtering and silence trimming, were applied uniformly across all samples to reduce variability introduced by recording environments. The class labels were assigned based on the original dataset annotations, which were derived from clinically validated sources. Although the dataset is curated, it relies on previously labeled data, and no additional manual relabeling was performed in this study. Therefore, the reliability of annotations depends on the quality and consistency of the original labeling process. Despite its usefulness, the dataset presents certain limitations. First, class imbalance exists, with COPD samples being more prevalent than bronchitis samples, which may influence model learning. Second, variations in recording devices and environmental noise conditions may introduce unintended biases. Third, the dataset may not fully represent diverse patient demographics, potentially affecting generalization to real-world clinical settings. To mitigate these issues, data augmentation techniques and multiple train-test splits were employed. However, future work will focus on validating the model on larger and more diverse datasets to further improve robustness and generalizability.

4.2. Performance Metrics

Sensitivity:

Represents the ability to correctly detect actual asthma cases.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (16)$$

Precision:

Indicates the proportion of correctly predicted positive cases among all predicted positives:

$$Precision = \frac{TP}{TP + FP} \quad (17)$$

F1 Score:

The F1-score, a weighted average of precision and recall, is a key metric for assessing a classification model's accuracy.

$$F1Score = \frac{2 * (Recall * Precision)}{Recall + Precision} \quad (18)$$

Specificity:

Specificity measures the percentage of true negatives a model accurately detects, indicating its ability to distinguish the negative class and reduce false positives.

$$Specificity = \frac{TN}{TN + FP} \quad (19)$$

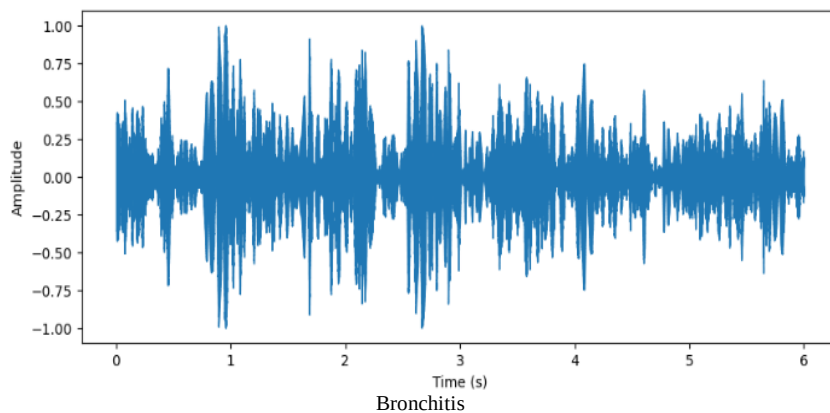
Accuracy:

This metric measures the percentage of correctly classified samples across all classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

Where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

The sample wave forms of the audio from the dataset are shown in Fig. 2. The performance evaluation of the proposed method based on various splits of training and testing percentages is shown in Table 1. The classification performance improves significantly across all metrics, including accuracy, specificity, sensitivity, and F1-score, as the training percentage increases. The model achieved an accuracy of 0.57, specificity of 0.57, sensitivity of 0.55, and an F1-score of 0.53 when the dataset was split equally for training and testing (50%-50%), and the performance showed a slight improvement, reaching an accuracy of 0.60 when the training percentage increased to 55%-45%. Performance continued to improve with each training data allocation increment; a noticeable increase in accuracy of 0.84 was noted at 70%-30%. With an accuracy of 0.98, specificity of 0.99, sensitivity of 0.99, and F1-score of 0.95, the best results were obtained at 80%-20%, suggesting a well-generalized model. Although the proposed model achieves high accuracy at an 80/20 train-test split, this result should be interpreted with caution due to the limited dataset size. Such high performance may partially reflect overfitting to the training distribution rather than true generalization. To mitigate this, regularization techniques and data augmentation were applied; however, these measures alone may not fully eliminate overfitting risks.



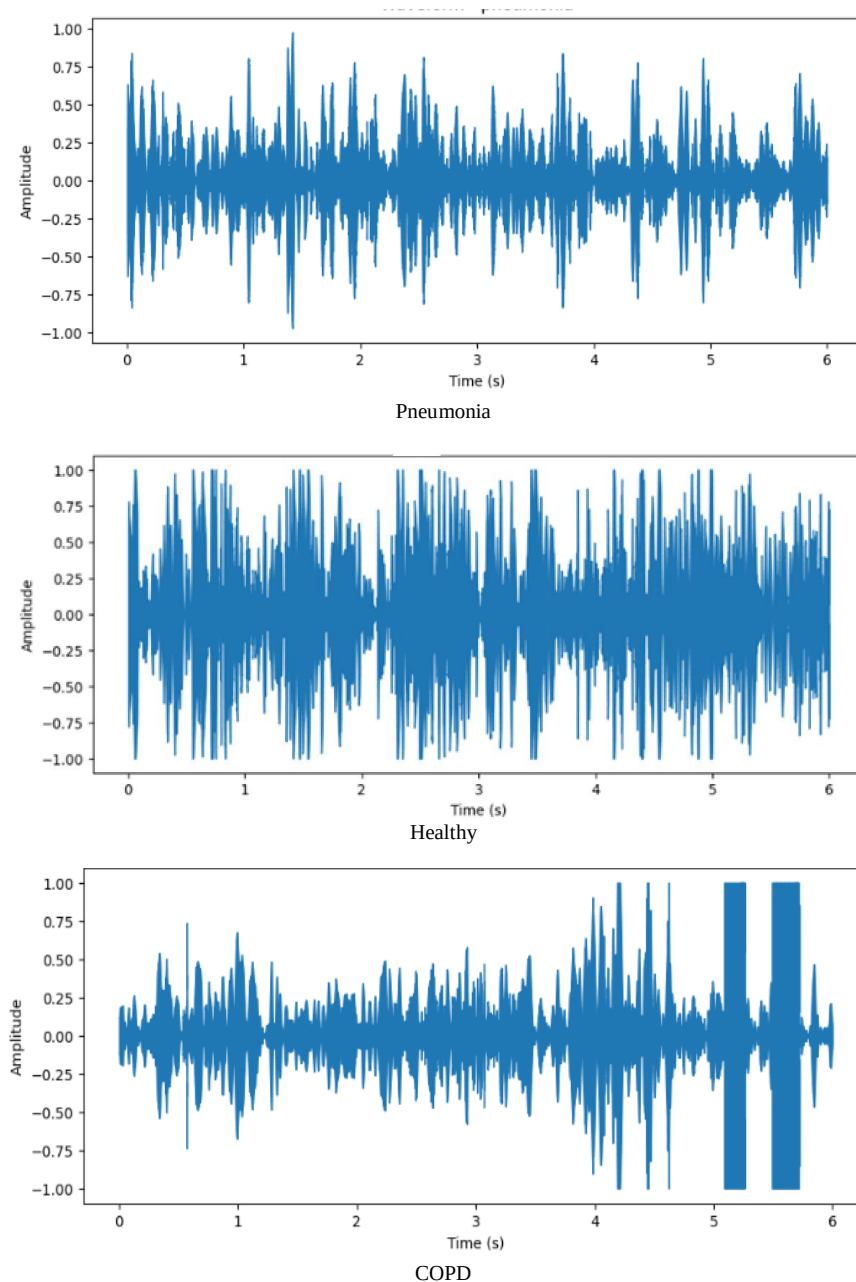
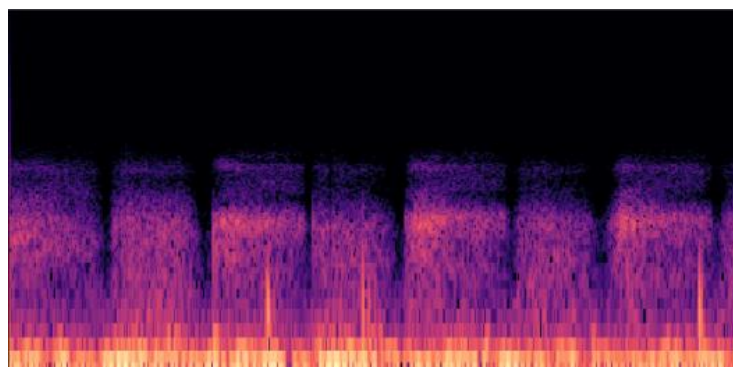


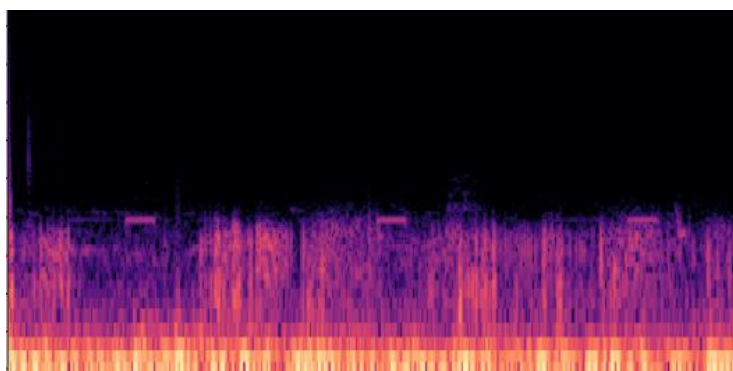
Fig. 2. Sample wave forms of the audio from the dataset.

The sample spectrogram pictures are displayed in Fig. 3. A performance comparison of the suggested method's single-scale feature extraction and multi-scale feature fusion is shown in Table 2. In every assessment criterion, the multi-scale technique performs noticeably better than the single-scale approach. The accuracy increases from 84.15% to 99.21%, indicating that multi-scale feature fusion is a useful technique for capturing more complex breathing patterns. Likewise, sensitivity rises from 75.64% to 99.56%, suggesting improved asthma-related feature recognition. Increasing the specificity from 90.47% to 100% guarantees a decreased false positive rate. Additionally, the accuracy increases from 88.24% to 98.48% and the F1-score improves from 84.39% to 98.64%, demonstrating the robustness of the suggested strategy in differentiating asthma from normal breath sounds. These results confirm that multi-scale feature fusion enhances classification performance by leveraging diverse feature representations. A comparison of the proposed method's performance against existing models, such as ResNet, CNN-based approaches, and SVM, is shown in Table 3. The proposed method's highest accuracy of 99.21% outperforms ResNet (98.79%), CNN models (88.6% and 89.3%), and SVM (99.10%); its sensitivity is 99.56%, outperforming ResNet (96.27%), CNN (86.47% and 90.13%), and SVM (98.50%); its specificity is 100%, which is comparable to ResNet and higher than CNN (86.62% and 91.57%) and SVM (98.72%); furthermore, the proposed model's F1-score of 98.64% and precision of 98.48% indicate a well-balanced classification performance. Although the proposed model achieves high performance metrics, including 99.21% accuracy and 100% specificity, these results may be influenced by dataset characteristics such as segment-level splitting

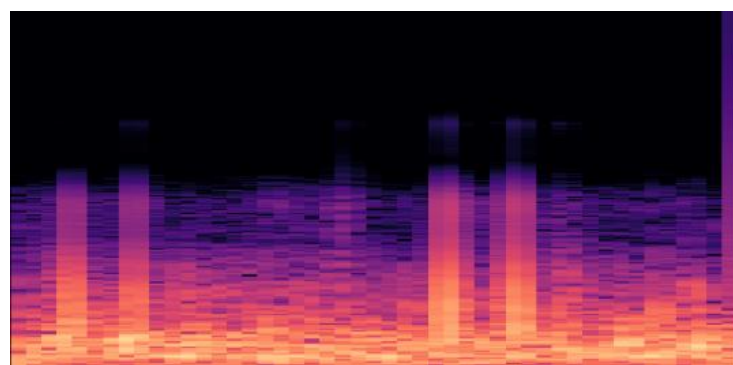
and potential similarity between training and testing samples. In particular, if multiple segments originate from the same recording or subject, this may introduce unintended data leakage, leading to optimistic evaluation results. Fig.4. shows the comparative performance of the proposed method.



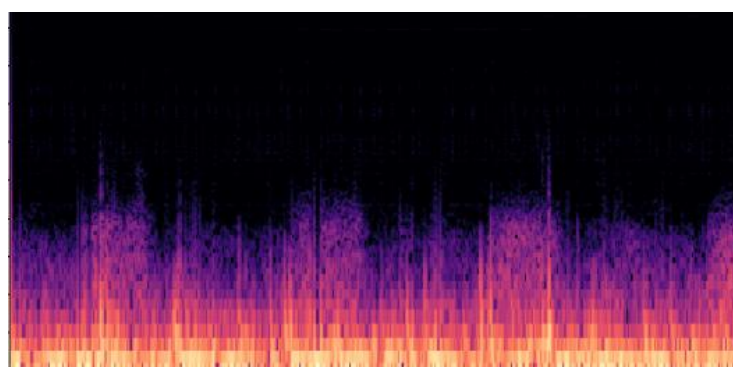
Bronchitis



Pneumonia



Asthma



Healthy

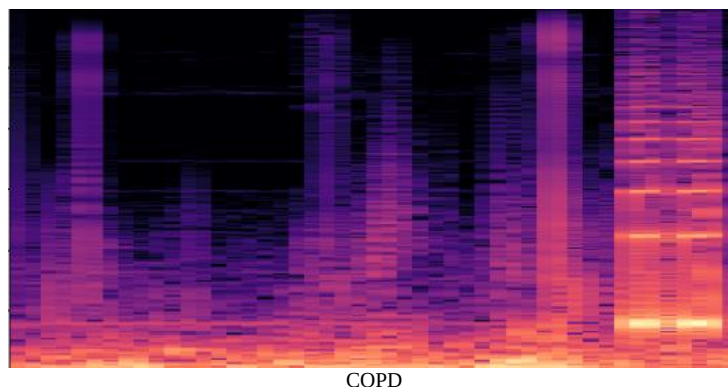


Fig. 3. Sample Spectrogram images.

Table 1. Performance of the proposed method based on training and testing percentage.

Training (%)	Testing (%)	Performance of classification			
		Accuracy	Specificity	Sensitivity	F1 score
50	50	0.57	0.57	0.55	0.53
55	45	0.60	0.52	0.64	0.62
60	40	0.65	0.67	0.65	0.63
65	35	0.71	0.73	0.75	0.70
70	30	0.84	0.84	0.81	0.83
75	25	0.90	0.95	0.92	0.89
80	20	0.98	0.99	0.99	0.95

Table 2. Performance comparison of the proposed method using single-scale and multi-scale feature extraction.

Performance	Single-scale feature extraction	Multi-scale feature fusion
Accuracy	84.15	99.21
Sensitivity	75.64	99.56
Specificity	90.47	100
F1- Score	84.39	98.64
Precision	88.24	98.48

Table 3. Comparative performance of the proposed method.

Performance	RESNET [7]	CNN [11]	CNN [12]	SVM [13]	This work
Accuracy	98.79	88.6	89.3	99.10	99.21
Sensitivity	96.27	86.47	90.13	98.50	99.56
Specificity	100	86.62	91.57	98.72	100
F1- Score	97.18	88.51	91.58	98.45	98.64
Precision	98.17	87.46	92.46	96.54	98.48

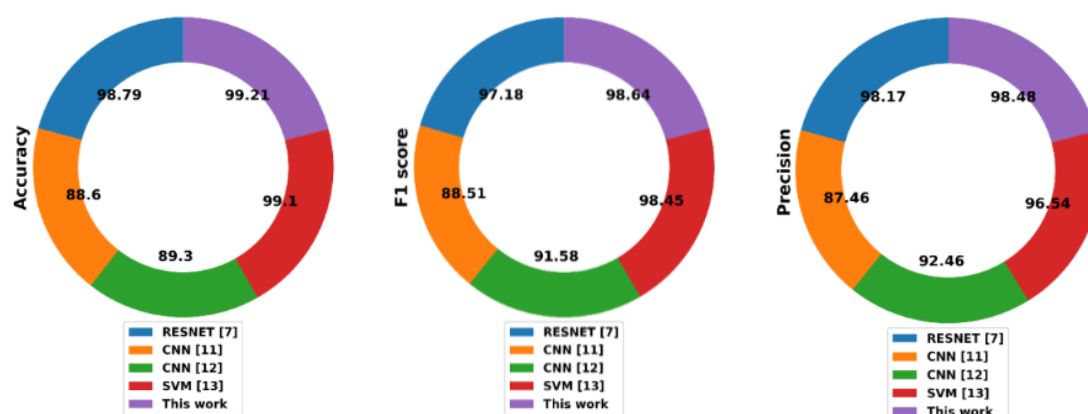


Fig. 4. Comparative performance of the proposed method

4.3. Future work

While the proposed model demonstrates high performance on the available dataset, the relatively small size and self-curated nature of the dataset may limit its generalizability to broader clinical populations. Respiratory sound characteristics can vary significantly across different age groups, recording devices, and environmental conditions.

Although data augmentation and preprocessing techniques were employed to mitigate these effects, the model's performance should be further validated on larger, standardized, and clinically diverse datasets. Future work will focus on cross-dataset evaluation using publicly available benchmarks and real-world clinical data to ensure robustness and wider applicability.

5. Conclusion

In this study, we proposed a multi-scale feature fusion framework based on EfficientNet for the detection of asthma and COPD using respiratory sound spectrograms. By integrating preprocessing techniques, STFT-based feature extraction, and spectrogram augmentation, the model is able to capture both local and global acoustic patterns, resulting in improved classification performance. Despite these promising results, several limitations must be acknowledged. First, the dataset used in this study is relatively small and self-curated, which may limit the generalizability of the model to diverse clinical populations. Second, the use of segment-level data splitting may introduce potential bias, leading to optimistic performance estimates. Third, although augmentation and regularization techniques were applied, the risk of overfitting cannot be entirely eliminated. Furthermore, real-world deployment of such models in clinical settings presents additional challenges, including variability in recording devices, environmental noise, and the need for standardized data acquisition protocols. Integration with clinical workflows and validation against large-scale, multi-center datasets are necessary before practical adoption. Future work will focus on addressing these limitations by incorporating subject-wise data splitting, evaluating the model on publicly available benchmark datasets, and conducting cross-dataset validation. Additionally, efforts will be made to develop lightweight and interpretable models suitable for real-time clinical applications.

All the Declarations and Statements

Author Contributions Statement

G. M. Karthik – Conceptualization, Methodology, and Supervision: Proposed research ideas, constructed the overall framework, and supervised project execution.

D. Lakshmi Padmaja – Data Curation and Software Implementation: Handled data acquisition, dataset preprocessing, and implementation of the research model.

Nageswara Rao Medikundu – Model Training, Validation, and Performance Evaluation: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods.

C. Sateesh Kumar Reddy – Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation.

Pichili Vijaya Bhaskar Reddy – Writing – Original Draft Preparation: Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.

Anoop V. – Writing – Review & Editing and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and coordinated project milestones and deadlines.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research received no external funding.

Data Availability Statement

This study analyzed publicly available datasets. The datasets and experimental results are available at:

<https://www.kaggle.com/datasets/mohammedtawfikmusaed/asthma-detection-dataset-version-2>

Accessed on: May 2026.

Ethical Declarations

This study does not involve human participants or animals.

Acknowledgments

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, the authors used Generative AI tools for language refinement, grammatical correction, and formatting support only. All generated content was carefully reviewed, validated, and edited by the authors. The authors take full responsibility for the content and integrity of the final manuscript.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
 CNN - Convolutional Neural Network
 COPD - Chronic Obstructive Pulmonary Disease
 DL - Deep Learning
 LSTM - Long Short-Term Memory
 ML - Machine Learning
 PFT - Pulmonary Function Test
 STFT - Short-Time Fourier Transform
 SVM - Support Vector Machine

References

- [1] A. Nair, H. Vadher, P. Patel, T. Vyas, C. Bhatt, and A. Bruno, "Lung sound disease detection using attention over pre-trained EfficientNet architecture," *Multimedia Tools and Applications*, pp. 1–35, 2024. doi:10.1007/s11042-024-20078-1
- [2] M. Habib, V. Vicente-Palacios, and P. García-Sánchez, "Bio-inspired optimization of feature selection and SVM tuning for voice disorders detection" *Knowledge-Based Systems*, vol. 310, pp.112950, 2025. doi:10.1016/j.knosys.2024.112950
- [3] N. Vodnala, P. S. Yarlagadda, M. Ch, and K. Sailaja, "Novel deep learning approaches to differentiate asthma and COPD based on cough sounds" in *Proceedings of the 2024 Parul International Conference on Engineering and Technology (PICET)*, IEEE, pp. 1–4, 2024. doi: 10.1109/PICET60765.2024.10716083.
- [4] P. Sahu, S. Kumar, and A. K. Behera, "DeepFusion: Early diagnosis of COPD, asthma, and pneumonia using lung sound analysis with a multimodal BiGRU network," *Computer Methods in Biomechanics and Biomedical Engineering*, pp. 1–14, 2025. doi:10.1080/10255842.2025.2511228.
- [5] P. Sahu, S. Jha, and S. Kumar, "Optimized 1D CNNs for enhanced early detection and accurate prediction of COPD and other pulmonary diseases," in *Proc. IEEE Region 10 Symp. (TENSYP)*, 2024, pp. 1–6. doi: 10.1109/TENSYP61132.2024.10752156.
- [6] P. Sahu, P. Prasad, V. P. Verma, and S. Kumar, "Deep learning framework for early diagnosis of COPD and respiratory diseases using lung sound analysis," in *Proc. Int. Conf. Big Data Analytics*, Cham: Springer Nature Switzerland, 2024, pp. 295–304. doi: 10.1007/978-3-031-81821-9_17.
- [7] H. Chen, X. Yuan, Z. Pei, M. Li, and J. Li, "Triple-classification of respiratory sounds using optimized S-transform and deep residual networks," *IEEE Access*, vol. 7, pp. 32845–32852, 2019. doi: 10.1109/ACCESS.2019.2903859.
- [8] M. A. Iqbal, K. Devarajan, and S. M. Ahmed, "An optimal asthma disease detection technique for voice signal using hybrid machine learning technique," *Concurrency Comput.: Pract. Exp.*, vol. 34, no. 11, p. e6856, 2022. doi: 10.1002/cpe.6856.
- [9] N. Zakaria, M. Fezari, A. Redjati, and K. Sundaraj, "Three ResNet deep learning architectures applied in pulmonary pathologies classification," in *Proc. Int. Conf. AI for Cyber Security Systems and Privacy (AI-CSP)*, 2021, pp. 1–8.
- [10] R. Karla and R. Yalavarthi, "A hybrid RNN-based deep learning model for lung cancer and COPD detection," *Eng., Technol. Appl. Sci. Res.*, vol. 14, no. 5, pp. 16847–16853, 2024. doi: 10.48084/etasr.8181.
- [11] V. Koshta, B. K. Singh, A. K. Behera, and T. G. Ranganath, "Fourier decomposition-based automated classification of healthy, COPD, and asthma using single-channel lung sounds," *IEEE Trans. Med. Robot. Bionics*, vol. 6, no. 3, pp. 1270–1284, 2024. doi: 10.1109/TMRB.2024.3408325.
- [12] A. Roy, B. Gyanchandani, A. Oza, and A. Singh, "TriSpectraKAN: A novel approach for COPD detection via lung sound analysis," *Scientific Reports*, vol. 15, no. 1, p. 6296, 2025. doi: 10.1038/s41598-024-82781-1
- [13] X. Zhang, J. Fu, W. Wang, and L. Yu, "Diagnosis of chronic obstructive pulmonary disease based on deep learning and auscultation lung sound," *Biomedical Signal Processing and Control*, vol. 112, p. 108438, 2026. doi: 10.1016/j.bspc.2025.108438.
- [14] S. D. Aaron et al., "Early diagnosis and treatment of COPD and asthma—a randomized, controlled trial," *N. Engl. J. Med.*, vol. 390, no. 22, pp. 2061–2073, 2024. doi: 10.1056/NEJMoa2401389
- [15] Y. Yan, L. van Bommel, F. M. E. Franssen, S. O. Simons, and V. Urovi, "Developing a multi-feature fusion model for exacerbation classification in asthma and COPD," *Comput. Methods Programs Biomed.*, vol. 268, p. 108796, 2025. doi: 10.1016/j.cmpb.2025.108796.
- [16] G. Shanthakumari, S. Gayathri, M. Parvathababu, J. Raamkumar, and E. Prajin, "Lung disease diagnosis through Mel spectrogram and acoustic feature analysis," in *Proc. Int. Conf. Computing and Communication Technologies (ICCCT)*, 2025, pp. 1–8. doi: 10.1109/ICCCT63501.2025.11019006.
- [17] S. Xu, R. C. Deo, O. Faust, P. D. Barua, J. Soar, and R. Acharya, "Automated lightweight model for asthma detection using respiratory and cough sound signals," *Diagnostics*, vol. 15, no. 9, p. 1155, 2025. doi:10.3390/diagnostics15091155

- [18] P. Sahu, S. Kumar, and A. K. Behera, "SoundNet: Leveraging deep learning for the severity classification of chronic obstructive pulmonary disease based on lung sound analysis," in Proc. IEEE Int. Conf. Electronics, Computing and Communication Technologies (CONECCT), 2024, pp. 1–6. doi: 10.1109/CONECCT62155.2024.10677193.
- [19] K.-R. Lee et al., "A wearable stethoscope for accurate real-time lung sound monitoring and automatic wheezing detection based on an AI algorithm," Engineering, 2025. doi: 10.1016/j.eng.2024.12.031
- [20] M. M. Ramyasri, M. Yoga, C. Harish, M. M. Rizwan, and K. Monisha, "Utilizing neural networks to analyze respiratory sounds for the detection of lung diseases," in Proc. Int. Conf. Advances in Computing and Communications (ICACC), 2024, pp. 1–6. doi: 10.1109/ICACC63692.2024.10845461.
- [21] B. Zhang, Y. Gu, D. Fu, L. Zhou, X. Wang, L. Guo, and J. Wang, "MSMF-DBNet: Dual-branch model for pulmonary disease recognition with multi-scale feature enhancement and multichannel fusion," Applied Soft Computing, p. 114847, 2026, doi: 10.1016/j.asoc.2026.114847.
- [22] W. N. A. Ismail, N. Ghazali, and R. Ngadiran, "Development of a robust deep learning network to distinguish COPD and asthma based on the augmentation of imbalanced respiratory sounds," in AIP Conf. Proc., vol. 3338, no. 1, p. 020009, 2025. doi: 10.1063/5.0295943.
- [23] M. L. Bennett, "Automated identification of respiratory anomalies from cough acoustics using spectrogram-driven deep learning," Frontiers Emerg. AI Mach. Learn., vol. 2, no. 1, pp. 8–13, 2025.
- [24] T. Albiges, "Artificial intelligence and signal analysis for COPD classification: Detecting and understanding of respiratory disease severities," Ph.D. dissertation, Bournemouth Univ., 2025.

Authors' Profiles



G. M. Karthik received the B.E. degree in Computer Science and Engineering from SACS MAVMM Engineering College, Madurai, India, and the M.E. degree in Computer Science and Engineering from PSNA College of Engineering and Technology, Dindigul, India, in 2003 and 2005, respectively. He obtained his Ph.D. degree in Information and Communication Engineering from Anna University, Chennai, India. He is currently working as an Associate Professor in the School of Computer Science Engineering (SCOPE), Vellore Institute of Technology, Vellore, India. He has more than 17 years of teaching experience. His research interests include Data Mining, Big Data Analytics, Web Mining, Time Series Data Analysis, Modelling, and Complexity Analysis.



D. Lakshmi Padmaja received her Master of Technology and Ph.D. degrees in Computer Science and Engineering from JNTU Hyderabad, India. She is currently working as an Associate Professor in the Department of Information Technology, Anurag University, Hyderabad, India. She has more than 23 years of teaching experience. Her research interests include Machine Learning, Software Engineering, and Algorithms. She has published research articles in reputed SCI, Scopus, and Web of Science indexed journals and conferences including IEEE and Springer publications.



Nageswara Rao Medikundu is currently working as an Associate Professor in the School of Mechanical and Civil Sciences at Koneru Lakshmaiah Education Foundation, Guntur, India. He has more than 13 years of teaching experience, 3 years of industry experience, and 8 years of research experience. He has published more than 100 research articles in reputed international journals and conferences. His research interests include Artificial Intelligence, Computing Techniques, Optimization Tools, and Advanced Communication Technologies.



C. Sateesh Kumar Reddy is currently working as a Senior Assistant Professor in the Department of Advanced Computer Science and Engineering at Vignan's Foundation for Science, Technology and Research, Guntur, India. He received his Ph.D. degree in Biomedical Signal Processing from VIT University, Chennai, India, in 2023. He has over thirteen years of teaching and research experience. His research interests include Biomedical Signal Processing, EEG Analysis, Machine Learning, Deep Learning, and Healthcare Analytics.



Pichili Vijaya Bhaskar Reddy received his Master's and Ph.D. degrees from the Department of Animal Science, University of Hyderabad, India. He completed his postdoctoral research at the University of Miami Miller School of Medicine, Florida, USA. He is currently working as an Assistant Professor in the Department of Life Science and Bioinformatics at Assam University Diphu Campus, India. His research interests include Neurobiology, Bioinformatics, and Biomedical Research. He has published more than 60 research articles in international journals.



Anoop V. is currently working as a Professor in the Department of Artificial Intelligence and Data Science at Jyothi Engineering College, Thrissur, India. He received his Ph.D. degree in Signal Processing from Visvesvaraya Technological University, Karnataka, India, in 2017. He completed his Master's degree in Communication Systems from Anna University, Chennai, India. He has more than 19 years of teaching experience. His research interests include Biomedical Image Processing and Speech Processing.

How to cite this paper:

G. M. Karthik, D. Lakshmi Padmaja, Nageswara Rao Medikundu, C. Sateesh Kumar Reddy, Pichili Vijaya Bhaskar Reddy, Anoop V., "Efficient ResNet-based Deep Learning Model for Asthma and COPD Detection Using Short-Time Fourier Transform Spectrograms", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.5, pp. 1-15, 2026. DOI:10.5815/ijigsp.2026.05.01