Emerging Science Innovation, vol. 7, 2026, pp. 01-13

Automated Lung Sound-Based Disease Classification Using a Multi-Model Pre-Trained CNN Framework

Nisha Palanisamy*, Vijayakumar Jeganathan

Department of Electronics and Instrumentation, Bharathiar University, Coimbatore, India

Received 15 April 2026; received in revised form 16 April 2026; accepted 20 April 2026

DOI: https://doi.org/10.46604/emsi.2026.16374

Abstract

The early identification of pulmonary diseases is essential for timely treatment and better patient care. This study presents a deep learning framework for automated classification of lung sound based diseases using nine pre-trained convolutional neural networks (CNNs). The lung sounds are filtered using a sixth-order Butterworth bandpass filter and converted into log-Mel spectrograms before training. Evaluations of the candidate models include to ensure the reliability and clinical relevance of the diagnostic results. EfficientNet-B0 achieves the highest accuracy of 99.20%, precision of 99.20%, sensitivity of 99.20%, specificity of 99.73%, and F1-score of 99.22%, followed by DenseNet201 (99.00%) and ResNet101 (98.90%). These results demonstrate that pre-trained CNNs can learn effective, discriminative features from lung sounds, leading to precise and efficient classification of respiratory diseases.

Keywords: lung sound classification, deep learning, convolutional neural networks, pre-trained models, computer-aided diagnosis

1. Introduction

Respiratory disorders remain a major global health challenge and are among the leading causes of morbidity and mortality worldwide. According to the World Health Organization, pulmonary diseases such as asthma, pneumonia, tuberculosis (TB), and chronic obstructive pulmonary disease (COPD) collectively cause more than three million deaths each year [1-2]. Early and accurate detection of these diseases is therefore essential for effective clinical intervention and improved patient outcomes.

Traditionally, lung auscultation with a stethoscope is a fundamental diagnostic tool in respiratory assessment. However, its diagnostic reliability depends heavily on clinical expertise, and subjective interpretation can lead to inconsistent results [3-4]. To address these limitations, computer-aided respiratory sound analysis (CARSA) has emerged as an effective approach that integrates digital stethoscopes with intelligent computational models for objective lung sound evaluation [5]. Such systems reduce observer bias, enable reproducible analysis, and allow large-scale screening in both clinical and telemedicine settings. With advances in deep learning and digital signal processing, computer-aided diagnostic (CAD) systems have achieved remarkable progress in audio and biomedical signal interpretation.

Recent research demonstrates that convolutional neural networks (CNNs) are particularly effective in capturing discriminative representations directly from time–frequency features such as spectrograms [6-7]. By transforming lung sounds into log-Mel spectrograms, CNNs can model both spectral and temporal variations that characterize respiratory abnormalities, offering an approach that closely resembles human auditory perception [8].

* Corresponding author. E-mail address: nisha.electronics@buc.edu.in

While several deep learning models have been applied to lung sound classification, there remains a limited comparative analysis of modern pre-trained architectures optimized for efficiency, generalization, and real-time application. Experimental results indicate that EfficientNet-B0 achieves the highest classification accuracy (99.20%), outperforming other models. Meanwhile, DenseNet201 and ResNet101 demonstrate strong generalization through residual and dense connections. Furthermore, lightweight models such as MobileNetV2, maintain competitive accuracy with reduced computational cost, highlighting their potential for embedded healthcare applications.

Joshi et al. [9] employed CNN models for cough sound classification, applying a fourth-order Butterworth high-pass filter to suppress noise and extract Mel-frequency cepstral coefficients (MFCCs). Their CNN architecture integrated max-pooling and dropout layers to minimize overfitting, achieving reliable differentiation of cough types through effective feature extraction and dimensionality reduction. Similarly, Oishee et al. [10] utilized the International Conference on Biomedical and Health Informatics (ICBHI) 2017 dataset and implemented multiple data augmentation methods such as noise addition, time shifting, stretching, and pitch modification. They compared feature extraction techniques, including MFCC, Mel-spectrogram, and chroma-short-time fourier transform (STFT) and proposed a hybrid CNN–long short term memory (LSTM) model for robust respiratory disease classification.

In another study, Zhang et al. [11] combined CNN and LSTM architectures to leverage both spatial and temporal features from lung sound spectrograms. Using the ICBHI 2017 dataset and extensive data augmentation, their dual-channel CNN–LSTM classifier demonstrated enhanced precision and stability in recognizing multiple lung sound categories. Crisdayanti et al. [12] proposed an attention-based deep neural network (DNN) for automated lung sound classification, employing STFT, Mel-spectrogram, and MFCC features. Their model incorporated knowledge propagation to enhance generalization and achieved high sensitivity in identifying abnormal lung sounds. Other works focused on signal-level preprocessing using filters such as the fourth-order Butterworth and wavelet transforms to convert time-domain signals into frequency-domain representations suitable for CNN-based classification [13-14].

Recent approaches also leverage advanced deep learning techniques, including residual learning [14], mixture-of-experts (MoE) [15], and attention-based mechanisms. These frameworks emphasize improved feature fusion, enhanced gradient propagation, and reduced data loss. Consequently, the achieve superior classification performance on benchmark datasets such as ICBHI 2017. For example, Lam Pham et al. [15] implemented a CNN–MoE framework, demonstrating that multi-expert gating improves robustness and discrimination of respiratory anomalies.

In summary, previous studies confirm that deep learning models outperform traditional machine learning methods for lung sound classification. Previous studies have successfully demonstrated the capability of classifying respiratory diseases using traditional machine learning algorithms—such as support vector machines (SVM), k-nearest neighbors (KNN), random forest, and decision trees. Typically, prior research relied heavily on using hand-crafted features obtained by manipulating the sound signals, such as: time-domain characteristics, frequency characteristics, and statistical characteristics. A prominent concern with traditional machine learning systems is their limited ability to extract complex biomedical relationships through the processing of the original continuous data of change.

However, challenges remain concerning data scarcity, noise robustness, and computational efficiency. The present work advances this field by conducting a comprehensive comparative analysis of nine pre-trained CNN architectures. This study integrates optimized preprocessing, data augmentation, and feature representation techniques to achieve state-of-the-art classification accuracy with high computational efficiency.

This study aims to fill this gap by presenting a deep learning framework for the automated classification of lung sounds using multiple pre-trained CNN architectures. Raw auscultation recordings are first filtered using a sixth-order Butterworth bandpass filter to remove irrelevant frequency components and then converted into log-Mel spectrograms. These spectrograms

serve as two-dimensional input representations for fine-tuned CNN models. Nine architectures—including AlexNet, VGG16, GoogleNet, ResNet50, ResNet101, DenseNet201, EfficientNet-B0, MobileNetV2, and SqueezeNet—are systematically evaluated for classifying four categories of lung sounds: normal, asthma, COPD, and pneumonia. The main contributions of this paper are as follows:

(1) A spectrogram-based deep learning framework is developed for automatic lung sound classification across multiple respiratory conditions.

(2) A comprehensive comparative evaluation of nine pre-trained CNN architectures is conducted to determine their effectiveness in pulmonary sound analysis.

(3) Advanced preprocessing and augmentation techniques are incorporated to improve signal quality and model robustness.

(4) The study identifies EfficientNet-B0 as the most effective model, combining high diagnostic accuracy with computational efficiency.

The remainder of this paper is organized as follows: Section 2 describes the proposed methodology, including data preparation, preprocessing, and CNN configuration. Section 3 presents the experimental setup, results, and discussion, and Section 4 concludes the paper with key findings and future directions

2. Methodology

The proposed framework leverages pre-trained CNN architectures for automated classification of abnormal lung sounds. The methodology comprises five primary stages: (i) dataset collection, (ii) pre-processing, (iii) data augmentation, (iv) feature representation using log-Mel spectrograms, and (v) model training and evaluation. The workflow of the proposed system is illustrated in. Fig. 1.

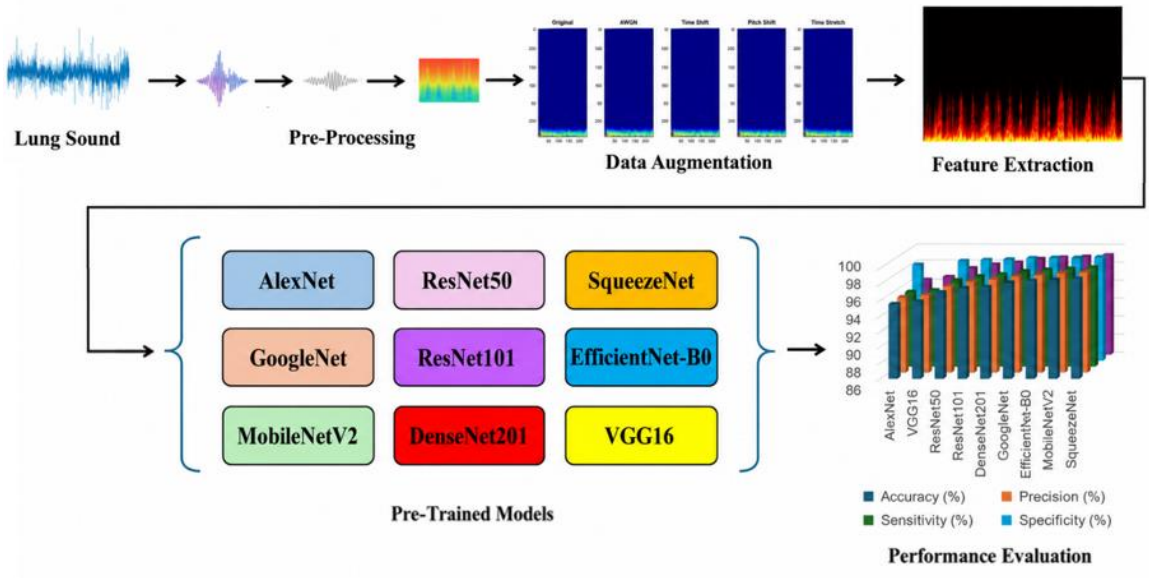


Fig. 1 Workflow of CNN architectures

The overall workflow of the proposed system is illustrated in Fig. 1. The process begins with the acquisition of lung sound recordings, followed by preprocessing using a Butterworth bandpass filter to remove noise and retain relevant frequency components. The filtered signals are then converted into log-Mel spectrograms, which serve as input features for the CNN models. Data augmentation techniques are applied to improve model generalization. Finally, multiple pre-trained CNN architectures are trained and evaluated using performance metrics such as accuracy, precision, sensitivity, specificity, and F1-score.

2.1 Dataset collection

The dataset used in this study consists of auscultation recordings captured from both healthy individuals and patients diagnosed with respiratory diseases such as asthma, COPD, and pneumonia. Table. 1 shows the detailed distribution of the lung sound samples used in this research.

Table 1 Distribution of lung sound samples across categories in own and online datasets

Dataset	Own dataset	Online dataset
Normal	65	582
Asthma	150	477
COPD	58	634
Pneumonia	63	440

All recordings are obtained using a high-fidelity electronic stethoscope equipped with a sensitive microphone to ensure signal clarity and reduce background interference. To maintain uniformity across samples, all audio signals are resampled to 4 kHz and saved in *.wav* format, preserving the clinically relevant frequency range of lung sounds [16-17]. This categorization ensures balanced representation across normal and abnormal classes and facilitates multi-class classification. To improve generalization and enhance dataset diversity for model training, a publicly available dataset is utilized. This particular dataset is not fully standardized clinically—meaning there is variability in in data acquisition—as consistent metadata (such as age or gender) is unavailable, and significant variability exists in class distribution.

2.2 Pre-processing

Raw lung sound recordings often contain unwanted noise caused by stethoscope friction, heart sound, environmental disturbances, or patient movement. To mitigate these artifacts, a sixth-order Butterworth bandpass filter with a passband of 50–2000 Hz was applied [18]. This range covers the essential spectral components of inspiratory and expiratory sounds while removing low-frequency and high-frequency interference. The filtering process preserves diagnostically relevant spectral information associated with asthma, COPD, and pneumonia, thereby enhancing signal quality for subsequent feature extraction. Normalization is then performed on the filtered signals to ensure each signal maintains a consistent amplitude and reduce variability between recordings. The use of normalized signals as input data improves quality and increases the robustness of subsequent feature extraction and classification processes.

2.3 Data augmentation

Deep learning models are highly sensitive to dataset size and variability. To improve generalization and reduce overfitting, systematic data augmentation is applied at both waveform and spectrogram levels. Ths techniques are implemented on lung sound signals to increase the robustness and generalization of the model.

First, additive white gaussian noise (AWGN) is added to the signals to simulate background noise or interference. The noise is characterized by a mean of zero and a standard deviation of 0.01. Second, time shifting is performed by shifting each signal circularly forward or backward within a range of 100 to 500 samples. This step accounts for differences in recordings start times relative to respiratory phases. Third, pitch shifting is employed, including both raising and lowering the pitch by two semitones (±2 semitones). This process simulates anatomical and age-related variations, as well as the potential effects of specific abnormalities on pitch. Finally, time stretching and compression are applied using time scaling factors of 0.9 and 1.1, allowing the system to model diverse respiratory rates while preserving their original pitch characteristics. By combining these augmentation techniques, the model robustness is enhanced by accommodating multiple variations of acoustic and physiological inputs [19-22].

Log-Mel spectrograms are subjected to both frequency and time masking to introduce feature-level variability, thereby encouraging more generalized temporal–spectral feature learning. In combination, these augmentation techniques expand the effective dataset size nearly elevenfold, substantially enhancing model robustness and diversity. The augmented dataset is subsequently divided into training, validation, and testing subsets using an 80:10:10 split. To ensure statistical reliability, fivefold cross-validation is also employed.

2.4 Feature representation

Lung sound recordings are converted from one-dimensional time-domain waveforms into two-dimensional log-Mel spectrograms to capture both temporal and spectral characteristics. This representation effectively models the nonlinear nature of human auditory perception, as the Mel scale closely approximates how humans perceive changes in frequency. The conversion from linear frequency f (Hz) to Mel frequency m is mathematically expressed as:

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

Each pre-processed signal is analysed using the short-time fourier transform (STFT), followed by Mel-scale filtering and logarithmic compression to produce the final log-Mel spectrogram. This representation enhances the visibility of subtle amplitude variations linked to pathological conditions, making it an ideal two-dimensional input for CNN-based classification. Examples of spectrograms corresponding to normal and abnormal lung sounds including asthma, COPD, and pneumonia.

2.5 CNN architectures employed

In this study, nine pre-trained CNN architectures are evaluated to determine their suitability for classifying lung sounds. The selected models include AlexNet, VGG16, ResNet50, ResNet101, DenseNet201, GoogleNet, EfficientNet-B0, MobileNetV2, and SqueezeNet. Through transfer learning, each architecture is adapted to the spectrogram dataset, resulting in improved convergence efficiency and lower computational cost. A concise overview of their architectural characteristics is summarized in Table 2**.** These deep learning architectures offer distinct merits in feature extraction and computational efficiency.

AlexNet incorporates hierarchical feature representations with an increase number of layers, supported by ReLU activations and dropout regularization to improve convergence and generalization. VGG16 employs a uniform stack of 3×3-sized convolutional filters throughout the entire network to develop multi-level feature representations. Both ResNet50 and ResNet101 utilize residual skip connections to improve the stability of deep networks and to mitigate the vanishing gradient issue, allowing training for stable and efficacious training of deeper networks.

DenseNet201 improves both gradient flow and parameter efficiencies through the use of massive connectivity and thus promotes the ability to use large variety of features across multiple layers. GoogleNet captures multi-scale features with inception modules while minimizing computational costs and keeping a rich representation structure. The design of EfficientNet-B0 utilizes the conceptof compound scaling—uniformly transforming network depth, width, and resolution—to achieve an exceptional balance between accuracy and computational efficiency. These diverse architectures allow for a comprehensive evaluation of practical performance in automated lung sound classification.

Finally, MobileNetV2 and SqueezeNet feature compact and efficient designs, making them ideal for real-time and embedded healthcare systems. These architectures achieve strong performance through the use of depthwise separable convolutions and fire modules, respectively.

Table 2 Summary of CNN architectures employed in the proposed framework

Network	Total layers	Parameters	Design strategies	Efficiency features
AlexNet	8 (5 conv + 3 FC)	~60 million	Max-pooling, ReLU activations, fully connected (FC) layers, and sequential convolution.	Reduces overfitting via dropout in FC layers; enables effective feature extraction with moderate depth.
VGG16	16 (13 conv + 3 FC)	~138 million	Hierarchical feature learning with uniform 3x3 convolutions stacked sequentially.	Captures fine-grained patterns through deep architecture; ensures stable training with sequential design.
ResNet50	50	~25 million	3x3 convolution layers and residual blocks with identity shortcut connections.	Prevents vanishing in deep networks; maintains performance via residual connections.
ResNet101	101	~44 million	Additional 3-layer blocks in an extended residual network.	Enables learning of intricate patterns; facilitates gradient flow through deep residual connections.
DenseNet201	201	~20 million	All layers have dense connectivity, meaning that all inputs from earlier layers are received by each layer.	Enhances gradient propagation and feature reuse; achieves high representational power with fewer parameters.
GoogLeNet	22	~4 million	Convolutions and concatenation on multiple scales in the inception modules.	Reduces computation through effective parameter use; captures features at various scales.
EfficientNet-B0	237 (scaled)	~5.3 million	Compound resolution, width, and depth scaling.	Achieves high accuracy with minimal parameters; remains computationally efficient for high-dimensional inputs.
MobileNetV2	155	~3.4 million	Inverted residual blocks and depth-wise separable convolutions.	Optimizes low-power deployment with a lightweight design; minimizes computation while maintaining expressive power.
SqueezeNet	15	~1.2 million	Fire modules with smaller input channels and 1x1 and 3x3 filters.	Accelerates execution with large activation maps; offers > 50× fewer parameters than AlexNet.

2.6 Training strategy

All CNN architectures are fine-tuned using the Adam optimizer with an initial learning rate of 0.001, a batch size of 16, and a maximum of 500 epochs. To improve training stability, batch normalization is applied after each convolutional block, while dropout layers (ranging from 0.2 to 0.6) are introduced to mitigate overfitting and enhance model generalization. The learning objective minimizes the categorical cross-entropy loss, which is mathematically expressed as:

$$L = -\sum_{i=1}^{c} y_i \log\left(\hat{y}_i\right) \tag{2}$$

where C is the number of classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability for the i-th class. Here, i is the class index. An early stopping criterion is applied to monitor validation loss, halting training once the model shows no meaningful improvement; this maintains efficiency and prevents overfitting. Among all evaluated architectures, EfficientNet-B0 achieves the highest classification accuracy, demonstrating the advantages of compound scaling and efficient feature extraction for biomedical audio signal analysis.

3. Results and Discussion

3.1. Experimental setup

All experiments are carried out in MATLAB R2022a on a workstation equipped with an Intel Core i5 processor (4.0 GHz), 16 GB RAM, and an NVIDIA GeForce RTX 4050 GPU. The study utilizes 7510 lung sound recordings divided into four categories: normal, asthma, COPD, and pneumonia. Nine pre-trained CNN architectures—including EfficientNet-B0, DenseNet201, ResNet101, ResNet50, MobileNetV2, GoogLeNet, VGG16, AlexNet, and SqueezeNet—are evaluated to determine the most effective model for multi-class lung sound classification.

Before model training, all recordings are pre-processed using a sixth-order Butterworth bandpass filter (50–2000 Hz) to remove background interference and retain clinically meaningful frequency components. The cleaned signals are then transformed into log-Mel spectrograms, serving as two-dimensional image inputs for CNN-based analysis. Each network is fine-tuned using the Adam optimizer with an initial learning rate of 0.001, a batch size of 16, and up to 500 epochs. Batch normalization and dropout are employed to stabilize learning and improve generalization. To ensure reliable evaluation, 10-fold (K=10) is employed. The main hyperparameters, including the dropout rate and the number of neurons in fully connected layers, are optimized through experimentation to balance classification accuracy, training stability, and computational cost.

3.2. Evaluation metrics

Model performance is assessed using standard classification metrics: accuracy (A), precision (P), sensitivity (S), and F1-score (F1), derived from true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics are mathematically defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision (P) measures how many of the predicted positive samples are truly positive:

$$precision = \frac{TP}{TP + FP} \tag{4}$$

Sensitivity (S), also known as Recall, quantifies the model's ability to identify genuine positive cases:

$$recall = \frac{TP}{TP + FN} \tag{5}$$

Specificity, often referred to as the true negative rate, measures the model's capability to correctly recognize normal or healthy cases. It reflects how effectively the system avoids false alarms by distinguishing non-diseased samples from abnormal ones:

$$specificity = \frac{TN}{TN + FP} \tag{6}$$

The F1 represents the harmonic mean of precision and sensitivity, offering a balanced measure of overall classification quality:

$$F1 - Score = \frac{precision \cdot recall}{precision + recall} \times 2 \tag{7}$$

In addition to these metrics, computational efficiency—including prediction speed (samples per second) and total execution time—is analyzed to assess the suitability of each architecture for real-time or portable diagnostic application.

3.3. Comparative performance of CNN architectures

Among the evaluated architectures, EfficientNet-B0 achieves the highest accuracy (99.20%), surpassing all other models. DenseNet201 and ResNet101 follow closely with accuracies of 99.00% and 98.90%, respectively, while ResNet50 reaches 98.70%. The lightweight MobileNetV2 attains 98.00%, offering an excellent trade-off between accuracy and computational efficiency.

Deeper models, such as DenseNet201 and ResNet101, benefit from improved gradient propagation and feature reuse, allowing them to maintain stable learning even in deeper layers. Compact architectures, such as MobileNetV2 and SqueezeNet, deliver competitive accuracy while maintaining low memory usage and computational requirements. These characteristics make them well-suited for real-time clinical deployment and portable diagnostic applications. Table 3 summarizes the performance, key hyperparameters, and parameter counts of all models, demonstrating the relationship between model complexity and diagnostic precision. A comparative summary of the CNN models is outlined in Table 4.

Table 3 Overview of CNN architectures, training settings, and performance

Model	Number of parameters in millions	Suggested hyperparameters	Accuracy (%)
SqueezeNet	1.2M	LR: 0.001, Batch: 16, Epochs: 500, 10-fold	95.80
AlexNet	61M	LR: 0.001, Batch: 16, Epochs: 500, 10-fold	96.10
VGG16	138M	LR: 0.001, Batch: 16, Epochs: 500, 10-fold	97.40
GoogLeNet	7M	LR: 0.001, Batch: 16, Epochs: 500, 10-fold	97.90
MobileNetV2	3.4M	LR: 0.001, Batch: 16, Epochs: 500, 10-fold	98
ResNet-50	25M	LR: 0.001, Batch: 16, Epochs: 500, 10-fold	98.70
ResNet-101	44M	LR: 0.001, Batch: 16, Epochs: 500, 10-fold	98.90
DenseNet-201	20M	LR: 0.001, Batch: 16, Epochs: 500, 10-fold	99
EfficientNet-B0	5.3M	LR: 0.001, Batch: 16, Epochs: 500, 10-fold	99.20

Table 4 presents the detailed performance metrics across all models

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
SqueezeNet	95.80	95.82	95.80	98.60	95.80
AlexNet	96.10	96.10	96.10	90.70	96.11
VGG16	97.40	97.39	97.40	99.13	97.39
GoogleNet	97.90	97.91	97.90	99.30	97.90
MobileNetV2	98.00	98.20	98.22	99.40	98.20
ResNet50	98.70	98.70	98.70	99.57	98.70
ResNet101	98.90	98.90	98.90	99.63	98.90
DenseNet201	99.00	99.10	99.00	99.67	99.00
EfficientNet-B0	99.20	99.20	99.20	99.73	99.22

3.4. Analysis of classification results

The classification results across all architectures reveal several important trends. Architectures such as ResNet and DenseNet, characterized by extensive connections and hierarchical layers, demonstrate superior sensitivity and specificity compared to simpler networks due to efficient gradient flow and multi-level feature extraction. Among all evaluated networks, EfficientNet-B0 use only 5.3 million parameters and performed nearly perfectly in classifying lung sounds, ignificantly outperforming VGG16, which possesses approximately 138 million parameter. This outcome highlights the efficacy of compound scaling and suggests that the excessive size of VGG16 is unnecessary for this specific task.

Lightweight networks, such as MobileNetV2, achieve a high level of classification accuracy (approximately 98%) with only 3.4 million parameters. Their computational efficiency demonstrates significant potential for developing real-time diagnostic systems deployed at the edge or in a cloud environment. Misclassification analysis indicates that most error occurrs as a result of overlapping acoustic signatures between asthma and COPD.

The convergence patterns illustrated in Figs. 2-3 indicate steady improvements in training accuracy and a consistent decline in validation loss, demonstrating stable learning behavior with minimal overfitting across most models. This research evaluates the performance of deep learning model that uses a log-Mel spectrogram to classify abnormal lung sounds. To achieve this, nine pre-trained CNNs are applied as part of a comprehensive evaluation process. The approach incorporating a a sixth order Butterworth bandpass filter and transforms the signals into a log-Mel spectrogram. This process improve the clinical relevance of these representation(s) by creating spectral-temporal features for model input during training.

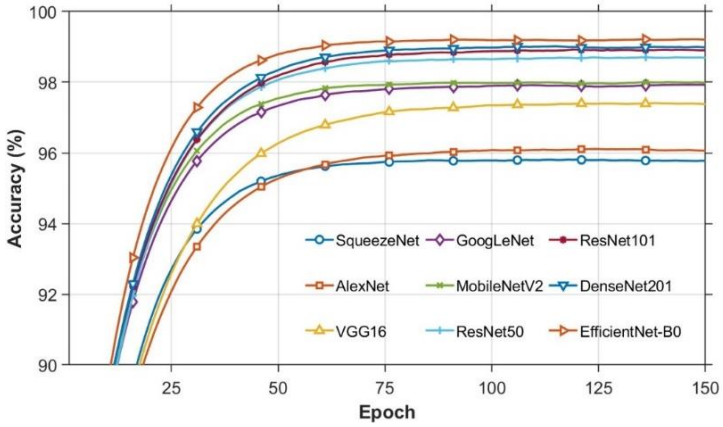


Fig. 2 Accuracy curves of CNN architectures on lung sound-based disease classification

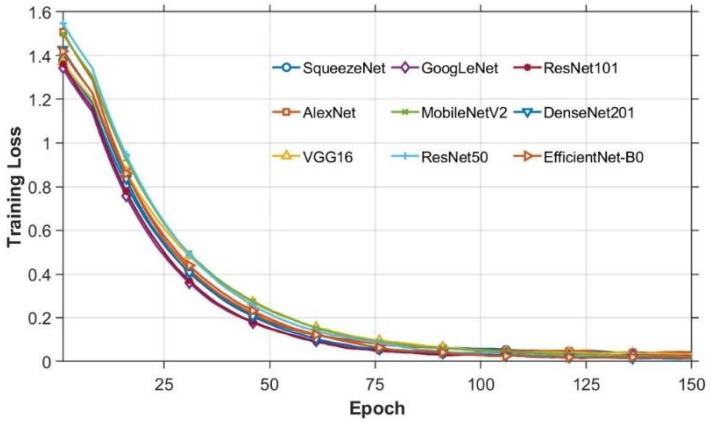


Fig. 3 Training loss curves of CNN architectures on lung sound-based disease classification

3.5. Quantitative esults

The superior performance of EfficientNet-B0 and DenseNet201 demonstrates the effectiveness of modern CNN scaling and dense connectivity in biomedical signal classification. The performance of CNN architectures on lung sound classification illustrated in Fig. 4.

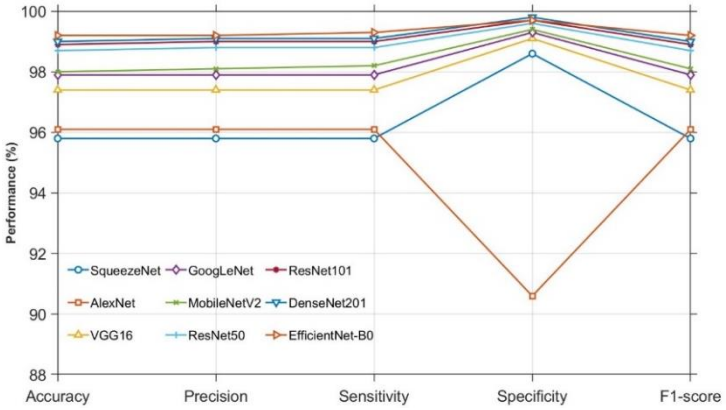


Fig. 4 The performance of CNN architectures on lung sound-based disease classification

3.6. Discussion

Fig. 5 shows the confusion matrix for the EfficientNet-B0 model. Confusion matrices for other models are omitted due to space constraints. The empirical findings confirm that spectrogram-based CNN architectures provide a highly effective solution for automated lung sound classification. EfficientNet-B0 achieves near-perfect diagnostic accuracy, reflecting its balanced design across network depth, width, and input resolution. The DenseNet201 and ResNet101 architectures also demonstrate strong performance, emphasizing the value of residual and dense feature propagation in distinguishing complex respiratory patterns.

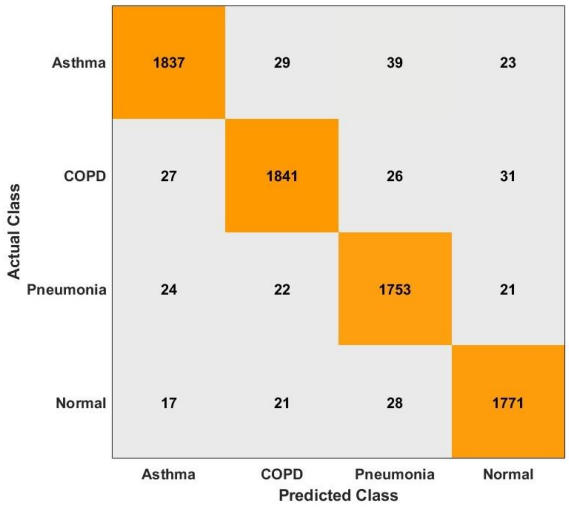


Fig. 5 EfficientNet-B0 confusion matrix

Table 5 Comparison of the proposed model with existing methods for lung sound classification

Methodology	Feature extraction	Dataset	Accuracy
Ensemble [7]	Mel spectrogram, MFCC, and Chromagram	ICBHI 2017	86.11%
CNN-LSTM [10]	MFCC, Mel Spectrogram, and Chroma STFT	ICBHI 2017	96%
Dual-channel CNN-LSTM [11]	MFCC	ICBHI 2017	99.01%
CNN [12]	STFT, Mel Spectrogram, and MFCC	Own dataset	87.55%
Bi-ResNet [13]	STFT and Wavelet Transform	ICBHI 2017	77.81%
CNN-LDA-RSE [14]	Spectrogram	ICBHI 2017	71.15%
CNN-GRU with Self-Attention [16]	Mel Spectrogram	ICBHI 2017	60.14%
RDLINet [21]	Mel Spectrogram	ICBHI 2017	99.6%
Proposed model (EfficientNet-B0)	log-Mel spectrogram	Own dataset	99.20%

Importantly, lightweight models such as MobileNetV2 and SqueezeNet achieved high accuracy with significantly reduced computational overhead. This trade-off positions them as ideal candidates for portable diagnostic tools**,** mobile health applications**,** and edge-AI deployments in telemedicine. Overall, the proposed framework establishes a reliable foundation for AI-driven respiratory diagnosis, demonstrating the potential of deep learning to augment clinical decision-making in pulmonary healthcare. Table 5 presents the comparison between existing and proposed work performance. These empirical results underscore three critical insights, as follows:

(1) Spectrogram representation provides an informative feature space for pathological lung sound analysis.

(2) Pre-trained CNNs can be effectively fine-tuned for medical audio data, eliminating the need for handcrafted features.

(3) Model selection must balance diagnostic accuracy with computational efficiency, depending on deployment context.

Recent research evaluates various methodologies for the classification of lung sounds. The results are determined on diverse datasets and demonstrate a wide range of performance across those studies. The traditional ensemble approaches produced an average accuracy of 86.11% on ICBHI 2017 dataset. More advanced deep recurrent architectures, such as CNN-

LSTM, yield improved results with an average accuracy of 96% on ICBHI 2017. Furthermore, adding an additional channel to create a dual-channel CNN-LSTM produces even more remarkable results, achieving an average accuracy of 99.01%, which highlights the advantages of multi-channel feature extraction.

The simple CNNs trained on lung sounds receive lower accuracies (e.g. 87.55% accuracy on self-collected datasets). Additionally, architectures such as Bi-ResNet and CNN-LDA-RSE produce significantly lower accuracies of 77.81% and 71.15%, respectively, which demonstrates that both architectures have limitations for lung sound classification. Models utilizing attention mechanisms, such as CNN+GRU+Self-attention, perform less than satisfactorily with only 60.14% accuracy on the ICBHI 2017 dataset.

The best performing model to date, RDLINet, produces an outstanding result of 99.6% accuracy on the ICBHI 2017 dataset. The proposed framework, based on EfficientNet-B0, delivers a strong 99.20% accuracy on self-collected datasets, confirming it is a competitive performer when compared to previous works. The classification task exhibits greater clarity and reliability due to the cleanliness and organization of the dataset, as well as the absence of extraneous noise. Consequently, this study serves as a strong benchmark for evaluation of high-precision, lightweight deep learning architectures in clinical diagnostics.

4. Conclusion

Each CNN model was fine-tuned and assessed using uniform evaluation protocols to evaluate its ability to classify respiratory sounds across diverse pathology categories. The key findings and implications of this research are summarized as follows:

(1) EfficientNet-B0 performed the highest overall accuracy with an overall accuracy of 99.20%, demonstrating the effectiveness of compound scaling in accurately classifying respiratory sounds.

(2) DenseNet201 and ResNet101 demonstrated consistently strong and reliable performance in classifying respiratory sounds, demonstrating advantages to using dense and residual connections for biomedical acoustic feature extraction.

(3) Lightweight models (i.e., MobileNetV2 and SqueezeNet) achieved comparable accuracy with very little computational burden and therefore would be appropriate for portable/real-time diagnostic applications.

(4) The combination of a Butterworth bandpass filter and log-Mel spectrogram representation greatly enhanced the model quality of the CNNs, allowing the CNN models to accurately capture the various pathologies associated with lung sounds.

(5) Validation of lightweight architectures for deployment in portable and embedded healthcare systems.

Future work will focus on utilizing larger datasets and developing CNN-transformer hybrid refine feature extraction. Furthermore, the implementation of embedded solutions and rigorous clinical validation will be prioritized to confirm the technology's reliability in real-world environments.

Conflicts of Interest

The authors declare no conflict of interest.

Statement of Ethical Approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Statement of Informed Consent

Informed consent was obtained from all individual participants included in the study.

References

[1] A. Yorgancioglu, N. Khaltaev, J. Bousquet, and C. Varghese, "The Global Alliance Against Chronic Respiratory Diseases: Journey So Far and Way Ahead," Chinese Medical Journal, vol. 133, no. 13, pp. 1513-1515, 2020.

[2] J. Torre-Cruza, F. Canadas-Quesada, R. Cortina-Parajon, J. Ranilla-Pastor, N. Ruiz-Reyes, P. Vera-Candeas, et al., "A Novel Data Augmentation Technique Based on Wheezing Physiological Modeling Applied to Asthma Severity Management in Respiratory Sounds," Computers in Biology and Medicine, vol. 196, part C, article no. 119010, 2025.

[3] R. Lang, Y. Fan, G. Liu, and G. Liu, "Analysis of Unlabeled Lung Sound Samples Using Semi-Supervised Convolutional Neural Networks," Applied Mathematics and Computation, vol. 411, article no. 126511, 2021.

[4] A. H. Sfayyih, N. Sulaiman, and A. H. Sabry, "A Review on Lung Disease Recognition by Acoustic Signal Analysis with Deep Learning Networks," Journal of Big Data, vol. 10, article no. 101, 2023.

[5] B. A. Tessema, H. D. Nemomssa, and G. L. Simegn, "Acquisition and Classification of Lung Sounds for Improving the Efficacy of Auscultation Diagnosis of Pulmonary Diseases," Medical Devices: Evidence and Research, vol. 15, pp. 89-102, 2022.

[6] G. Perera and H. C. Pathmakumara, "Advanced Deep Learning Techniques for Lung Sound Classification: Binary, Multi-Class, and Ensemble Approach," Proceedings of 2025 7th International Conference on Software Engineering and Computer Science (CSECS), IEEE Xplore, pp. 21-23, 2025.

[7] Q. Chen, W. Zhang, X. Tian, X. Zhang, S. Chen, and W. Lei, "Automatic Heart and Lung Sounds Classification Using Convolutional Neural Networks," Proceedings of 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), IEEE Xplore, pp. 13-16, 2017.

[8] A. Joshi, T. Kumar, and C. Tiwari, "Enhanced Exploration of Chronic Cough Using Improved Convolutional Neural Networks and Remote Monitoring Harnessing Internet of Things (IoT)," Materials Today: Proceedings, vol.46, part 15, pp. 6465-6473, 2021.

[9] T. T. Oishee, J. Anjom, U. Mohammed, and M. I. A. Hossain, "Leveraging Deep Edge Intelligence for Real-Time Respiratory Disease Detection," Clinical eHealth, vol. 7, pp. 207-220, 2024.

[10] 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," Biomedical Signal Processing and Control, vol. 94, article no. 106257, 2024.

[11] I. A. P. A. Crisdayanti, S. W. Nam, S. K. Jung, and S.-E. Kim, "Attention Feature Fusion Network via Knowledge Propagation for Automated Respiratory Sound Classification," IEEE Open Journal of Engineering in Medicine and Biology, vol. 5, pp. 383-392, 2024.

[12] C. Wu, N. Ye, and J. Jiang, "Classification and Recognition of Lung Sounds Based on Improved Bi-ResNet Model," IEEE Access, vol. 12, pp. 73079-73094, 2024.

[13] F. Demir, A. M. Ismael, and A. Sengur, "Classification of Lung Sounds with CNN Model Using Parallel Pooling Structure," IEEE Access, vol. 8, pp. 105376-105383, 2020.

[14] L. Pham, H. Phan, R. Palaniappan, A. Mertins, and I. McLoughlin, "CNN-MoE Based Framework for Classification of Respiratory Anomalies and Lung Disease Detection," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 8, pp. 2938-2947, 2021.

[15] P. Bhushan, M. S. Fahad, S. Agrawal, K. S. D. Kamesh, G. Singh, P. Tripathi, et al., "A Self-Attention Based Hybrid CNN-LSTM Architecture for Respiratory Sound Classification," GMSARN International Journal, vol 18, pp. 54–61, 2024,

[16] T. Aptekarev, V. Sokolovsky, E. Furman, N. Kalinina, and G. Furman, "Application of Deep Learning for Bronchial Asthma Diagnostics Using Respiratory Sound Recordings," PeerJ Computer Science, vol. 9, article no. e1173, 2023.

[17] J. Li, J. Yuan, H. Wang, S. Liu, Q. Guo, Y. Ma, et al., "LungAttn: Advanced Lung Sound Classification Using Attention Mechanism with Dual TQWT and Triple STFT Spectrogram," Physiological Measurement, vol. 42, article no. 10, 2021.

[18] J. Acharya and A. Basu, "Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning," IEEE Transactions on Biomedical Circuits and Systems, vol. 14, no. 3, pp. 535-544, 2020.

[19] B. TaghiBeyglou, A. Assadi, A. Elwali, and A. Yadollahi, "TRespNET: A Dual-Route Exploratory CNN Model for Pediatric Adventitious Respiratory Sound Identification," Biomedical Signal Processing and Control, vol. 93, article no. 106170, 2024.

[20] A. Roy and U. Satija, "RDLINet: A Novel Lightweight Inception Network for Respiratory Disease Classification Using Lung Sounds," IEEE Transactions on Instrumentation and Measurement, vol. 72, article no. 4008813, 2023.

[21] S. Kumar, A. V. Shvetsov, and S. H. Alsamhi, "Empowering Remote Healthcare with Federated Learning for Early Diagnosis of Pulmonary Disease," IEEE Internet of Things Journal, vol. 12, no. 13, pp. 23288-23296, 2025.

Copyright© by the authors. Licensee TAETI, Taiwan. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC) license (https://creativecommons.org/licenses/by-nc/4.0/).