JSCE (Journal of System and Computer Engineering)
Vol.7, No.4, October 2026, pp. 374-384
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i4.2837

Research Article Open Access (CC–BY-SA) ■

Optimizing a CNN–LSTM Model for Respiratory Disease Classification from Lung Sounds

Bahrum Herdania Bahar [1,a]; Sri Kurniyan Sari [1,b]; Radus Batau [1,c]; Firman Aziz [2,d,*]; Jeffry [3,e]

[1] *Universitas Indonesia Timur, Rappocini Raya Street No.171-173, Makassar 90222, Indonesia*
[2] *Universitas Pancasakti, Jl. Andi Mangerangi No. 73, Makassar 90121, Indonesia*
[3] *Institut Teknologi Bacharuddin Jusuf Habibie, Jl. Balaikota No.1, Parepare 91122, Indonesia*
[a] *bahrumbahar616@gmail.com*; [b] *srikurniyans@gmail.com*; [c] *radus@uit.ac.id*; [d] *firman.aziz@unpacti.ac.id*; [e] *jeffry@ith.ac.id*
* *Corresponding author*

Abstract

Respiratory disease classification from lung sounds remains challenging because of overlapping acoustic characteristics, imbalanced class distributions, and limited subject diversity. This study aims to optimize a Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) model for multiclass respiratory disease classification while evaluating strategies to address class imbalance. The study used a lung sound dataset initially containing 336 recordings from 112 subjects. After excluding unsuitable records and consolidating low-frequency categories, 324 recordings from 108 subjects across seven classes were retained. The recordings were normalized, standardized to 20 s, and converted into 64-bin Mel-spectrograms with 160 time steps. CNN, LSTM, and CNN–LSTM architectures were compared using a subject-level holdout evaluation. The CNN–LSTM achieved an accuracy of 34.85%, precision of 24.61%, recall of 34.85%, weighted F1-score of 28.22%, and macro-F1 of 14.07%. A separate five-fold Group Cross-Validation experiment was conducted to assess subject-independent performance and evaluate imbalance-handling strategies. The baseline CNN–LSTM achieved a mean accuracy of 35.87 ± 5.43% and a mean macro-F1 of 10.26 ± 3.50%. Class-Balanced Cross-Entropy increased macro-F1 to 12.16%, an improvement of 1.90 percentage points, while accuracy decreased to 27.52%. Data augmentation, including noise addition, gain variation, and time shifting, was applied to training data. Although Class-Balanced Cross-Entropy improved macro-F1, the results indicate that balanced recognition across all classes remains limited. The findings highlight the challenges posed by class imbalance and the small number of independent subjects. Further evaluation using larger and more diverse datasets is required to improve generalization and establish the model's suitability for clinical applications.

Keywords—*CNN–LSTM, Lung Sounds, Respiratory Disease Classification, Mel-Spectrogram, Class Imbalance.*

1. Introduction

Respiratory diseases represent a major global health concern and require accurate and timely diagnosis to support appropriate clinical management. Lung auscultation is one of the commonly used approaches for assessing respiratory conditions, as abnormal lung sounds can provide important information regarding the presence of airway and pulmonary abnormalities (Fraiwan et al., 2021). However, conventional auscultation largely depends on the experience

and auditory skills of clinicians, which may lead to variations in interpretation, particularly when abnormal sounds are subtle or recordings are affected by environmental noise and recording conditions (Kim et al., 2021). The development of digital healthcare technologies and artificial intelligence (AI) provides an opportunity to develop automated systems capable of analyzing lung sounds consistently and supporting respiratory disease assessment.

Respiratory sound analysis (RSA) has evolved from conventional signal-processing techniques toward machine learning and deep learning approaches. One of the major challenges in lung sound analysis is the complex temporal and frequency characteristics of respiratory signals. Consequently, time-frequency representations such as spectrograms, Mel-spectrograms, and Mel-frequency cepstral coefficients (MFCCs) have been widely explored to transform acoustic signals into representations that can be effectively processed by deep learning models (Serbes et al., 2018). These representations preserve relevant information regarding the distribution of acoustic energy across time and frequency, allowing computational models to learn patterns associated with normal and abnormal respiratory sounds. Recent studies have demonstrated that convolutional neural networks (CNNs), recurrent neural networks, and hybrid deep learning architectures remain promising approaches for automated respiratory sound classification (Yu et al., 2025).

Convolutional Neural Networks (CNNs) have demonstrated strong capabilities in extracting local patterns and spatial representations from time-frequency representations such as spectrograms. In lung sound analysis, CNNs can learn discriminative frequency and temporal patterns associated with respiratory abnormalities, including wheezes and crackles, without relying entirely on manually engineered features (Zhaoping Wang & Sun, 2024). Nevertheless, CNNs are primarily designed to capture local patterns and do not explicitly model long-term temporal dependencies. This limitation is particularly relevant to respiratory sound analysis because lung sounds occur sequentially throughout respiratory cycles, and clinically meaningful characteristics may depend on temporal relationships between consecutive signal segments. Long Short-Term Memory (LSTM), a specialized recurrent neural network architecture, is capable of learning temporal dependencies and can therefore complement CNN-based feature extraction in sequential respiratory sound analysis (Hochreiter & Schmidhuber, 1997).

The integration of CNN and LSTM has consequently become an attractive approach for respiratory sound classification. CNNs can extract high-level representations from spectrogram-based inputs, while LSTM networks can model the temporal dependencies among the extracted feature sequences. Fraiwan *et al.* demonstrated that a combination of CNN and LSTM can be employed to recognize pulmonary diseases from lung sound recordings (Bardou et al., 2018; Fraiwan et al., 2021; Jung et al., 2021). Other research has further demonstrated the potential of hybrid CNN-LSTM architectures for capturing both spatial and temporal characteristics of respiratory signals, while the use of focal loss has been investigated to address class imbalance during model training (Petmezas et al., 2022). More recent work has also explored Mel-spectral representations and dual-channel CNN-LSTM architectures to integrate complementary spatial and temporal information for lung sound classification (Zhang et al., 2024).

Despite these developments, several methodological challenges remain in the development of reliable respiratory sound classification systems. Class imbalance, variations in recording characteristics, background noise, differences in signal duration, and the limited number of subjects can affect the generalization capability of deep learning models (Zijie Wang & Wang, 2022; Yu et al., 2025). Furthermore, reported performance across studies is not always directly comparable because of differences in preprocessing procedures, data augmentation strategies, class distributions, dataset partitioning, and evaluation protocols. Consequently, a high classification accuracy alone may not adequately demonstrate the robustness and generalization capability of a respiratory disease classification model.

A previous study by Wahyudi *et al.* developed a CNN-BiLSTM-based approach for respiratory disease classification using lung sound visualization and considered four categories: asthma, bronchitis, tuberculosis, and normal (Wahyudi et al., 2025). The study demonstrated the

potential of combining CNN-based feature extraction with BiLSTM-based temporal modeling for respiratory disease classification and proposed its potential application in telemedicine. However, the previous work primarily focused on demonstrating the feasibility of the CNN-BiLSTM architecture, while systematic optimization of a CNN-LSTM architecture and its training parameters was not the main focus. In addition, issues related to class imbalance, data augmentation, model generalization, and rigorous evaluation remain important considerations in respiratory sound classification research (Petmezas et al., 2022; Yu et al., 2025).

Therefore, further research is needed to develop an optimized CNN-LSTM model that systematically investigates the influence of model architecture and training parameters on respiratory disease classification performance. Optimization can involve several important parameters, including the number of convolutional filters, kernel size, number of LSTM units, dropout rate, learning rate, batch size, and optimization algorithm. Such an approach is relevant because recent studies have demonstrated that integrating appropriate signal representations, deep feature extraction, temporal modeling, and model optimization can improve the effectiveness of automated respiratory disease classification (Ayappan et al., 2026; Zhang et al., 2024). In addition, data augmentation and class-balancing strategies can be incorporated to improve the ability of the model to learn from underrepresented respiratory disease categories (Petmezas et al., 2022; Zijie Wang & Wang, 2022).

This study aims to develop and optimize a **CNN-LSTM model for respiratory disease classification from lung sounds** using a publicly available lung sound dataset. The proposed framework focuses on a complete processing pipeline consisting of audio preprocessing, time-frequency feature representation, CNN-based feature extraction, LSTM-based temporal modeling, and hyperparameter optimization. To provide a comprehensive assessment, the optimized CNN-LSTM model will be compared with baseline CNN, LSTM, and conventional CNN-LSTM models. Model performance will be evaluated using accuracy, precision, recall, F1-score, and confusion matrix, with additional evaluation metrics considered where appropriate. The experimental design is intended to provide a systematic assessment of whether model optimization can improve respiratory disease classification performance compared with non-optimized baseline architectures.

The main contribution of this study is the development of an **optimized CNN-LSTM framework for respiratory disease classification from lung sounds**, with optimization treated as a central methodological component rather than simply applying a conventional CNN-LSTM architecture. The study further provides a comparative evaluation of CNN, LSTM, CNN-LSTM, and optimized CNN-LSTM models to identify the contribution of spatial feature extraction, temporal modeling, and hyperparameter optimization to classification performance. By employing a lung sound dataset that differs from the dataset used in the previous CNN-BiLSTM study, this research also provides an additional evaluation setting for assessing the applicability and generalization of hybrid deep learning approaches to respiratory sound classification. The findings are expected to contribute to the development of AI-based respiratory sound analysis and provide a methodological foundation for future intelligent clinical decision-support and telemedicine applications.

2. Method

2.1 Dataset and Preprocessing

This study uses the **Lung Dataset** obtained from Kaggle as the primary data source for developing a respiratory disease classification model based on lung sounds. The dataset contains lung sound recordings collected from subjects with different respiratory conditions. At the initial stage, all recordings were examined to determine their audio format, sampling rate, duration, number of channels, class distribution, and subject identifiers when available. This examination was conducted to ensure data quality and identify recordings that were unsuitable for subsequent experiments.

Audio preprocessing was performed to standardize the characteristics of the input signals. Stereo recordings were converted to mono, and all audio signals were resampled using a consistent sampling rate. Amplitude normalization was subsequently applied to reduce variations in recording intensity. The recordings were then segmented into fixed-duration audio segments to ensure consistent input dimensions for the deep learning models. Padding was applied when the signal duration was shorter than the predefined segment length.

To obtain a representation suitable for lung sound analysis, the audio signals were transformed into **Mel-spectrograms** using Short-Time Fourier Transform (STFT) and a Mel filter bank. The resulting representation was used as the primary input to the CNN because it preserves the time-frequency characteristics of respiratory sounds. MFCC features may also be investigated as an additional representation to assess the contribution of complementary spectral characteristics to model performance.

Data augmentation was applied only to the training data to increase sample variability and reduce overfitting. The augmentation procedures included controlled noise addition, gain variation, and time shifting. If substantial class imbalance was identified, class weighting was incorporated during model training so that underrepresented classes could contribute more appropriately to the optimization process.

The dataset was partitioned at the **subject level** rather than randomly at the audio-file level. An initial split of 70% for training, 15% for validation, and 15% for testing was employed. All recordings originating from the same subject were assigned to the same subset. This strategy was adopted to minimize data leakage and provide a more realistic estimation of model generalization to previously unseen subjects.

2.2 CNN–LSTM Model Development

A hybrid CNN-LSTM model was developed to exploit both spatial and temporal characteristics of lung sound signals. The CNN was employed as a feature extractor to learn local patterns from the Mel-spectrogram representation, while the LSTM was used to model temporal dependencies among the features extracted by the CNN.

The CNN architecture consists of convolutional layers followed by ReLU activation functions, pooling layers, and dropout. The convolutional layers extract discriminative patterns from the time-frequency representation, while pooling reduces the feature dimensionality and retains the most informative characteristics. The resulting feature maps are subsequently reshaped into sequential feature representations before being passed to the LSTM layer.

The LSTM component learns temporal dependencies within the sequence of CNN-extracted features. The LSTM output is then passed through fully connected layers followed by the classification layer. For multiclass classification, a softmax activation function is employed at the output layer to generate the probability of each respiratory disease class.

Three baseline models are also developed for comparison: CNN, LSTM, and conventional CNN-LSTM without optimization. The CNN baseline evaluates the capability of convolutional feature extraction from the time-frequency representation, whereas the LSTM baseline evaluates temporal modeling capability. The conventional CNN-LSTM serves as the hybrid baseline before hyperparameter optimization. This comparative design allows the contribution of each modeling component to be evaluated systematically.

2.3 Hyperparameter Optimization and Experimental Scenarios

Hyperparameter optimization represents the main methodological component of this study. The objective is to identify an optimal CNN-LSTM configuration that provides the best validation performance while minimizing overfitting. The hyperparameters considered include the number of convolutional filters, kernel size, number of LSTM units, dropout rate, learning rate, batch size, optimizer, and number of training epochs.

The hyperparameter search is conducted using a systematic search strategy, such as grid search or random search, depending on the number of parameter combinations and available computational resources. Each configuration is trained using the training set and evaluated on the validation set. F1-score is considered one of the primary criteria for selecting the optimal configuration because the dataset may contain imbalanced class distributions. Early stopping is applied to terminate training when validation performance fails to improve for a predefined number of consecutive epochs.

The experiments are conducted in several stages. First, CNN is evaluated as the baseline model, followed by LSTM and conventional CNN-LSTM. The CNN-LSTM model is subsequently evaluated with data augmentation and class weighting when required. Finally, the best-performing hyperparameter configuration is selected as the **optimized CNN-LSTM** model. All experiments use the same dataset partitioning and preprocessing procedures to ensure that performance differences can be attributed primarily to the model architecture and optimization strategy.

2.4 Model Evaluation

Model performance is evaluated using the test set, which is not involved in model training or hyperparameter optimization. The primary evaluation metrics include **accuracy, precision, recall, and F1-score**. Accuracy measures the proportion of correctly classified samples, while precision and recall assess the correctness and completeness of predictions for each respiratory disease class.

F1-score is used as an important complementary metric because it combines precision and recall and is particularly useful when class distributions are imbalanced. For multiclass classification, both macro-average and weighted-average precision, recall, and F1-score are reported. Macro-average assigns equal importance to each class, whereas weighted-average accounts for the number of samples in each class.

In addition to numerical performance metrics, a **confusion matrix** is generated to examine the distribution of correct and incorrect predictions across respiratory disease classes. When appropriate, ROC-AUC is also calculated as an additional measure of the model's discriminative capability. The optimized CNN-LSTM is subsequently compared with the CNN, LSTM, and conventional CNN-LSTM baselines to determine whether hyperparameter optimization provides a measurable improvement in respiratory disease classification performance.

Overall, the proposed methodology is designed to determine whether systematic hyperparameter optimization can improve the performance of CNN-LSTM for respiratory disease classification from lung sounds. The comparison between baseline and optimized models also enables the contributions of CNN-based feature extraction, LSTM-based temporal modeling, and hyperparameter optimization to be examined systematically.

3. Results And Discussion

3.1 Dataset Characteristics and Preprocessing Results

The dataset initially consisted of **336 lung sound recordings from 112 subjects**. Each subject contributed multiple recordings with different auscultation characteristics. The initial examination of the annotation data indicated an imbalanced distribution across respiratory disease categories, with Normal and Asthma representing the largest classes, while Lung Fibrosis, Pneumonia, and several other conditions contained substantially fewer recordings.

To establish a consistent single-disease classification setting, subjects with combined diagnoses were excluded from the primary experiment. In addition, Bronchitis and Pleural Effusion were merged into an **Other** category because the number of subjects in each individual category was insufficient to support reliable multiclass learning. Following this selection process, the final experimental dataset consisted of **324 recordings from 108 subjects**,

representing seven classes: Normal, Asthma, Heart Failure, COPD, Pneumonia, Other, and Lung Fibrosis.

Each audio recording had a sampling rate of **4,000 Hz**, with durations ranging from 5 to 30 s. To obtain a consistent input representation, the signals were normalized and standardized to 20 s using padding or center cropping. The processed signals were then converted into **Mel-spectrograms with 64 Mel bins**. The resulting representations were normalized and resized to 160 time steps before being provided to the deep learning models.

The Mel-spectrogram representation provides a two-dimensional representation of the respiratory signal in the time-frequency domain. This representation enables the CNN component to extract local spectral patterns while preserving temporal information that can subsequently be modeled by the LSTM component.

An important characteristic of the experimental design was the use of **subject-level data partitioning**. Of the 108 subjects, 73 were assigned to the training set, 13 to validation, and 22 to testing, corresponding to 219, 39, and 66 recordings, respectively. Consequently, recordings originating from the same subject were not simultaneously assigned to training and testing subsets. This procedure provides a stricter evaluation of model generalization to previously unseen subjects.

3.2 Baseline Model Comparison

The initial experiment compared three architectures: CNN, LSTM, and CNN-LSTM. All models employed the same preprocessing procedure and dataset partitioning to ensure that performance differences were primarily attributable to the model architecture.

Table 1. Performance Comparison of Baseline Models

Model	Accuracy	Precision	Recall	F1-score	Macro-F1	F1-Score
CNN	33.33%	32.87%	33.33%	26.53%	14.83%	71.8%
LSTM	25.76%	14.77%	25.76%	18.66%	8.85%	77.5%
CNN-LSTM	34.85%	24.61%	34.85%	28.22%	14.07%	86.0%

The CNN-LSTM model achieved an accuracy of **34.85%** and an F1-score of **28.22%**, slightly outperforming the CNN model, which obtained an accuracy of 33.33% and an F1-score of 26.53%. The LSTM model produced the lowest performance, with an accuracy of 25.76% and an F1-score of 18.66%. These results are consistent with the experimental findings reported in the source material.

The relatively better performance of CNN-LSTM compared with LSTM indicates that extracting discriminative time-frequency features before temporal modeling is beneficial for lung sound classification. The CNN component learns localized patterns from the Mel-spectrogram, while the LSTM processes the resulting feature sequence to capture temporal dependencies.

However, the improvement of CNN-LSTM over CNN was relatively small. This finding suggests that model architecture alone does not represent the primary limitation of the current dataset. Instead, the limited number of independent subjects, substantial class imbalance, and acoustic similarities between respiratory conditions may restrict the ability of the model to learn robust decision boundaries.

The distinction between the number of recordings and the number of independent subjects is particularly important. Although 324 recordings were available for the experiment, they originated from only 108 subjects. Therefore, the effective number of independent observations available for evaluating generalization is substantially smaller than the number of audio files.

3.3 CNN-LSTM Optimization

The CNN-LSTM architecture was further optimized by evaluating different configurations of convolutional filters, LSTM units, dropout, learning rate, and weight decay. Model selection was performed based on validation performance, while the test set was retained for final evaluation. The selected configuration is presented in Table 2.

Table 2. Selected CNN-LSTM Configuration

Parameter	Configuration
Mel bins	64
Time steps	160
CNN filters	96
LSTM units	64
Dropout	0.30
Learning rate	0.0005
Weight decay	0.0001
Optimizer	Adam
Loss function	Cross-Entropy
Early stopping	Applied

The selected configuration provides a balance between feature extraction capacity and temporal modeling complexity. The use of 96 convolutional filters increases the representational capacity of the CNN component, while 64 LSTM units provide sufficient temporal modeling capacity without introducing excessive model complexity for a relatively small subject-level dataset.

A dropout rate of 0.30 was incorporated to reduce overfitting, while a learning rate of 0.0005 was used to promote more gradual optimization compared with the initial learning rate of 0.001.

Importantly, the optimization results indicate that increasing model complexity does not necessarily result in substantial performance improvements. This finding suggests that the primary bottleneck is not solely associated with hyperparameter selection but is also related to the limited number of subjects and the imbalanced distribution of disease categories.

3.4 Confusion Matrix Analysis

The confusion matrix revealed that the CNN-LSTM model performed substantially better on the Normal and Asthma categories than on the minority disease categories.

Table 3. Confusion Matrix of CNN-LSTM

Actual / Predicted	Normal	Asthma	Heart Failure	COPD	Pneumonia	Other	Lung Fibrosis
Normal	14	7	0	0	0	0	0
Asthma	8	8	2	0	0	0	0
Heart Failure	4	7	1	0	0	0	0
COPD	5	1	0	0	0	0	0
Pneumonia	1	2	0	0	0	0	0
Other	0	0	3	0	0	0	0
Lung Fibrosis	1	2	0	0	0	0	0

The model correctly classified **14 of 21 Normal recordings**, corresponding to a recall of 66.7%. For Asthma, 8 of 18 recordings were correctly identified, resulting in a recall of 44.4%. In contrast, the model failed to correctly classify the COPD, Pneumonia, Other, and Lung Fibrosis samples in the test set.

The resulting macro-F1 score was only **14.07%**, compared with a weighted F1-score of **28.22%**. This difference indicates that the model's overall performance was strongly influenced by the larger classes, particularly Normal and Asthma.

Therefore, accuracy alone does not adequately describe the classification capability of the model. The confusion matrix and macro-F1 score reveal that the model has not yet achieved balanced recognition across the seven respiratory disease categories.

3.5 Class-Level Performance

A more detailed evaluation of the CNN-LSTM model was performed using class-level precision, recall, and F1-score.

Table 4. Class-Level Performance of CNN-LSTM

Class	Precision	Recall	F1-score	Support
Normal	42.4%	66.7%	51.9%	21
Asthma	29.6%	44.4%	35.6%	18
Heart Failure	16.7%	8.3%	11.1%	12
COPD	0.0%	0.0%	0.0%	6
Pneumonia	0.0%	0.0%	0.0%	3
Other	0.0%	0.0%	0.0%	3
Lung Fibrosis	0.0%	0.0%	0.0%	3

The highest class-level performance was obtained for Normal, with an F1-score of **51.9%**, followed by Asthma with an F1-score of **35.6%**. This indicates that the model was better able to learn the acoustic characteristics of these two relatively well-represented categories.

The performance of Heart Failure was substantially lower, with a recall of only 8.3%. Most of the Heart Failure samples were incorrectly classified as Normal or Asthma. The minority classes showed even greater difficulties, with no correctly classified test samples for COPD, Pneumonia, Other, or Lung Fibrosis.

These results should not be interpreted solely as a limitation of the CNN-LSTM architecture. The number of independent subjects in these categories was very small. The dataset contained only four subjects with Lung Fibrosis, five with Pneumonia, and nine with COPD in the single-disease classification setting. Such limited representation makes it difficult for the model to learn disease-specific patterns that generalize to unseen subjects.

3.6 Discussion

The experimental results demonstrate that CNN-LSTM provides slightly better performance than the individual CNN and LSTM architectures. CNN-LSTM achieved an accuracy of **34.85%**, compared with 33.33% for CNN and 25.76% for LSTM. This improvement suggests that combining time-frequency feature extraction with temporal sequence modeling is beneficial for respiratory sound classification.

Nevertheless, the relatively small improvement over CNN indicates that the main challenge is not simply the selection of the neural network architecture. The dataset contains only 108 independent subjects after filtering, despite having 324 audio recordings. Subject-level partitioning, which was deliberately employed to prevent data leakage, further reduces the amount of independent information available for model training.

This finding highlights an important methodological consideration in lung sound classification. Randomly splitting individual audio recordings may produce artificially optimistic results when multiple recordings originate from the same subject. By enforcing subject-level separation, the present experiment provides a more conservative but more meaningful estimate of model generalization to previously unseen subjects.

From an architectural perspective, CNN extracts localized patterns from the Mel-spectrogram, whereas LSTM models temporal relationships within the extracted feature sequence. However, the results demonstrate that combining these two components is insufficient to overcome severe class imbalance and limited subject diversity. Therefore, model optimization should be accompanied by strategies addressing the data distribution itself.

The difference between weighted F1-score (**28.22%**) and macro-F1 (**14.07%**) is particularly informative. Weighted F1 is strongly affected by the larger classes, whereas macro-F1 assigns equal importance to every class. The considerably lower macro-F1 therefore indicates that the model remains weak in recognizing minority disease categories.

These findings provide an important extension of previous research using CNN-BiLSTM approaches for respiratory disease classification. Rather than focusing solely on increasing overall classification accuracy, the present study emphasizes subject-level evaluation and the effect of hyperparameter optimization under a highly imbalanced respiratory sound dataset. This provides a more rigorous perspective on the generalization capability of CNN-LSTM models.

The results further indicate that **model optimization alone is insufficient to obtain reliable performance across all disease categories**. Future improvements should therefore consider more comprehensive strategies, including realistic audio augmentation, focal or class-balanced loss functions, audio transfer learning, ensemble approaches, and larger subject-level respiratory sound datasets.

The limited number of minority-class subjects is particularly important. With only a small number of subjects representing Pneumonia and Lung Fibrosis, multiclass classification across seven categories becomes highly challenging. Increasing the number of independent subjects, rather than merely increasing the number of recordings, should therefore be prioritized in future research.

Overall, the experiments demonstrate that **CNN-LSTM performs better than LSTM and marginally better than CNN under the current experimental setting**, but its performance remains insufficient for clinical application. The findings emphasize that effective respiratory sound classification requires not only an appropriate deep learning architecture but also representative data, balanced class distributions, and rigorous subject-level validation.

3.7 Summary of Experimental Findings

Table 5. Summary of Main Experimental Results

Finding	Result
Initial recordings	336
Initial subjects	112
Final recordings	324
Final subjects	108
Training	219 recordings / 73 subjects
Validation	39 recordings / 13 subjects
Testing	66 recordings / 22 subjects
CNN accuracy	33.33%
LSTM accuracy	25.76%
CNN-LSTM accuracy	34.85%
CNN-LSTM precision	24.61%
CNN-LSTM recall	34.85%
CNN-LSTM F1-score	28.22%
Macro-F1	14.07%
Best-performing class	Normal (F1 = 51.9%)

The overall findings indicate that the primary limitation of the current experimental setting is **not solely the CNN-LSTM architecture but the combination of severe class imbalance, limited independent subjects, and substantial variability in respiratory sound characteristics**. Consequently, the optimization strategy should be directed not only toward architectural hyperparameters but also toward improving class representation and subject-level generalization.

4. Conclusions

This study evaluated a Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) model for multiclass respiratory disease classification using lung sound recordings. After data screening and category consolidation, the final dataset comprised 324 recordings from 108 subjects across seven classes. The recordings were preprocessed and represented as 64-bin Mel-spectrograms with 160 time steps. In the subject-level holdout evaluation, the CNN–LSTM achieved an accuracy of 34.85%, precision of 24.61%, recall of 34.85%, a weighted F1-score of 28.22%, and a macro-F1 score of 14.07%. These results indicate that the model's classification performance remained limited, particularly for underrepresented disease categories. A separate five-fold Group Cross-Validation experiment was conducted to assess subject-independent performance and investigate class-imbalance handling. The baseline CNN–LSTM achieved a mean accuracy of 35.87 ± 5.43% and a mean macro-F1 score of 10.26 ± 3.50%. The Class-Balanced Cross-Entropy strategy increased the macro-F1 score to 12.16%, representing an improvement of 1.90 percentage points, although accuracy decreased to 27.52%. This trade-off suggests that improving recognition across classes does not necessarily increase overall accuracy. Data augmentation using noise addition, gain variation, and time shifting was applied to the training data to support model learning while preventing augmentation-related data leakage into evaluation data. Overall, the findings demonstrate that respiratory disease classification from lung sounds remains challenging when the dataset contains imbalanced classes and a limited number of independent subjects. Future work should evaluate these strategies on larger and more diverse subject populations, investigate more effective methods for minority-class recognition, and report class-level performance consistently across validation folds. Further external validation is also required before the model can be considered suitable for clinical decision-support applications.

Acknowledgements

The authors would like to express their sincere gratitude for the financial assistance and institutional support provided by the **Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia** through the **Program Hibah Penelitian Dosen Pemula Tahun 2026**.

References

Ayappan, G., Sumathi, S., Mani, V., & Jebarani, P. E. (2026). Novel deep learning-based optimization framework for the classification of respiratory diseases using lung sound analysis. *Scientific Reports 2026 16:1*, *16*(1), 19018-. https://doi.org/10.1038/s41598-026-44936-0

Bardou, D., Zhang, K., & Ahmad, S. M. (2018). Lung sounds classification using convolutional neural networks. *Artificial Intelligence in Medicine*, *88*, 58–69. https://doi.org/10.1016/J.ARTMED.2018.04.008

Fraiwan, M., Fraiwan, L., Alkhodari, M., & Hassanin, O. (2021). Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory. *Journal of Ambient Intelligence and Humanized Computing 2021 13:10*, *13*(10), 4759–4771. https://doi.org/10.1007/S12652-021-03184-Y

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Jung, S. Y., Liao, C. H., Wu, Y. S., Yuan, S. M., & Sun, C. T. (2021). Efficiently Classifying Lung Sounds through Depthwise Separable CNN Models with Fused STFT and MFCC Features. *Diagnostics 2021, Vol. 11, Page 732*, *11*(4), 732. https://doi.org/10.3390/DIAGNOSTICS11040732

Kim, Y., Hyon, Y. K., Jung, S. S., Lee, S., Yoo, G., Chung, C., & Ha, T. (2021). Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. *Scientific Reports*, *11*(1), 1–11. https://doi.org/10.1038/S41598-021-96724-

7;SUBJMETA=139,228,692,700;KWRD=DIAGNOSIS,HEALTH+SERVICES

Petmezas, G., Cheimariotis, G. A., Stefanopoulos, L., Rocha, B., Paiva, R. P., Katsaggelos, A. K., & Maglaveras, N. (2022). Automated Lung Sound Classification Using a Hybrid CNN-LSTM Network and Focal Loss Function. *Sensors 2022, Vol. 22, Page 1232*, *22*(3), 1232. https://doi.org/10.3390/S22031232

Serbes, G., Ulukaya, S., & Kahya, Y. P. (2018). An automated lung sound preprocessing and classification system based onspectral analysis methods. *IFMBE Proceedings*, *66*, 45–49. https://doi.org/10.1007/978-981-10-7419-6_8/SAVE-RESEARCH

Wahyudi, A. E., Batau, R., Aziz, F., & Jeffry, J. (2025). A Deep Learning Approach to Respiratory Disease Classification Using Lung Sound Visualization for Telemedicine Applications. *Journal of System and Computer Engineering*, *6*(4), 306–311. https://doi.org/10.61628/JSCE.V6I4.2144

Wang, Zhaoping, & Sun, Z. (2024). Performance evaluation of lung sounds classification using deep learning under variable parameters. *EURASIP Journal on Advances in Signal Processing 2024 2024:1*, *2024*(1), 51-. https://doi.org/10.1186/S13634-024-01148-W

Wang, Zijie, & Wang, Z. (2022). A DOMAIN TRANSFER BASED DATA AUGMENTATION METHOD FOR AUTOMATED RESPIRATORY CLASSIFICATION. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2022-May*, 9017–9021. https://doi.org/10.1109/ICASSP43922.2022.9746941

Yu, S., Yu, J., Chen, L., Zhu, B., Liang, X., Xie, Y., & Sun, Q. (2025). Advances and Challenges in Respiratory Sound Analysis: A Technique Review Based on the ICBHI2017 Database. *Electronics 2025, Vol. 14, Page 2794*, *14*(14), 2794. https://doi.org/10.3390/ELECTRONICS14142794

Zhang, Y., Huang, Q., Sun, W., Chen, F., Lin, D., & Chen, F. (2024). Research on lung sound classification model based on dual-channel CNN-LSTM algorithm. *Biomedical Signal Processing and Control*, *94*, 106257. https://doi.org/10.1016/J.BSPC.2024.106257