Deep Learning for Audio-Based Respiratory Disease Classification Using Clinical Sound Data

Cherylene Callista Reksohartono
Data Science Program
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
cherylene.reksohartono@binus.ac.id

Crysantha Monica Lim
Data Science Program
Computer Science Department
School Of Computer Science
Bina Nusantara University
Jakarta, Indonesia
crysantha.lim@binus.ac.id

Alexander Agung Santoso Gunawan
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
aagung@binus.edu

Jeffrey Junior Tedjasulaksana
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
jeffrey.t@binus.edu

***Abstract* — Respiratory diseases, including asthma, Chronic Obstructive Pulmonary Disease (COPD), and pneumonia, are common global health concerns that contribute to morbidity and mortality worldwide. These conditions often present with an initial symptom such as a cough. Early and accurate diagnosis is important for effective management and improved patient outcomes. Traditional diagnostic methods, however, are often limited by subjectivity and external noise interference, emphasizing the need for innovative approaches. This research investigates the use of hybrid deep learning models for classifying respiratory diseases using clinical sound data. There were 8 classes included in the data including healthy with seven diseases; COPD, URTI, Bronchiectasis, Bronchiolitis, Pneumonia, LRTI, and Asthma. Data was simplified into only 2 classes; COPD and non-COPD (which includes healthy and the other 6 diseases). Utilizing audio recordings from the Respiratory Sound Database, this study applies data preprocessing, feature extraction, and a classification model for 2 classes (COPD and non-COPD). The study compares three hybrid architectures—CNN + RNN, CNN + LSTM, and CNN + SVM—to determine the optimal model. Metrics like F1-score, recall, accuracy, and precision are used in evaluating models. The results showed an overall great performance in CNN + LSTM with an accuracy of 88%. CNN + SVM also showed good capability in classifying the data with an accuracy of 85%. However, CNN + RNN performed quite poorly due to the incapability in capturing and remembering the patterns of the audio data. This research contributes to the development of AI-driven diagnostic tools as an early warning to respiratory diseases.**

Keywords—respiratory disease classification, deep learning, respiratory sound analysis, feature extraction, data augmentation

I. Introduction

Asthma, pneumonia, and Chronic Obstructive Pulmonary Disease (COPD) remain significant public health issues, contributing to high morbidity and mortality. Early diagnosis is crucial, but around 70%–80% of COPD cases go undiagnosed, with delayed detection leading to reduced life expectancy [1][2]. Conventional methods like stethoscope auscultation depend on clinician expertise and are affected by environmental noise, making them less reliable [3][4]. The development of digital health tools, such as digital stethoscopes and machine learning, enables automatic classification of respiratory sounds [5][6]. Respiratory sounds, including crackles and wheezes, indicate various lung conditions, and deep learning approaches like Convolutional Neural Networks (CNNs) have been effective in identifying discriminative acoustic features using MFCCs and spectrograms [4][5][7]. However, most studies focus on standalone models, and few have systematically compared hybrid architectures like CNN-RNN, CNN-LSTM, or CNN-SVM, limiting our understanding of their practical advantages in COPD screening.

This study adopts CNN as the primary architecture and explores its integration with Support Vector Machines (SVMs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. This hybrid modeling approach is designed to assess whether combining CNN's spatial learning capabilities with the sequential or decision-boundary strengths of other architectures can further improve classification performance in distinguishing COPD and non-COPD cases. To enhance the precision of differentiating between healthy and pathological lung sounds, these models make use of both convolutional and sequential learning [7]. Model robustness in noisy situations can be improved through the use of advanced feature extraction techniques, which have been the focus of recent studies [3][5].

II. Literature Review

A. COPD

COPD is characterized by persistent respiratory symptoms and irreversible airflow limitation, including chronic cough, sputum production, and dyspnea. These, along with fatigue and mental health factors, can affect speech and voice quality, and nighttime symptoms like wheezing or snoring disturb sleep [8]. Smoking is the primary cause, with pollution, aging, and genetics as contributing factors. Unlike asthma, COPD is progressive, non-reversible, and marked by exacerbations, comorbidities, and specific risk profiles [9].

B. Hybrid CNN for audio classification

Kong et al. (2020) found that CNN-based models trained on large audio datasets outperform traditional approaches by learning relevant audio features. [10]. CNNs are excellent at feature learning, which means they can extract and learn only the right features from the input data while attempting to eliminate the irrelevant information [11].

The process begins with data preparation, where audio is transformed into 2D log-mel spectrograms by dividing it into short, non-overlapping frames. Using features based on auditory perception models instead of traditional features improves classification [12]. These spectrograms are used as input for CNN architectures, such as DNNs, AlexNet, VGG, Inception V3, and ResNet-50, with input and output layers adjusted to fit the spectrogram size and number of classes. The models are trained using the Adam optimizer and cross-entropy loss. Evaluation uses Area Under the Curve (AUC) and mean Average Precision (mAP), with segment-level predictions averaged to generate an overall classification for each recording [13].

C. Previous Works

1) CNN-based Audio Classification

CNN-based audio classification has been widely applied for environmental and speaker recognition tasks. Massoudi et al. [14] developed a CNN model using mel-spectrograms derived from MFCCs to classify urban sounds across 10 categories. Their sequential CNN architecture, comprising Conv2D, MaxPooling2D, and Dense layers with dropout, achieved 91% accuracy, though real-world generalizability was limited by noise and variability. Similarly, Ashar et al. [15] proposed a hybrid speaker identification model combining CNN and MFCC features. Using a dataset of 60 Urdu-speaking individuals, they compared three models: CNN with spectrogram input, DNN with MFCC features, and a fusion model combining both via chi-square-based feature selection. All audio samples underwent preprocessing—including noise reduction, silence trimming, mono conversion, and resampling to 16 kHz—before being converted into spectrograms and MFCCs. The hybrid approach yielded the best results with 87.5% accuracy and 91% precision, highlighting the benefits of feature integration for robust audio classification.

2) Lung Disease Audio Classification

Several studies have investigated automated classification of lung sounds using both traditional and deep learning approaches. Srivastava et al. proposed a CNN-based system specifically for COPD detection using respiratory audio input, achieving an ICBHI score of 93%—the highest among comparable studies—although their model was not generalized to other respiratory conditions [16]. Brunese et al. adopted a two-stage classification scheme to identify and differentiate types of lung diseases using audio features, comparing Logistic Regression, SVM, Neural Networks, and kNN, with the neural network model achieving the best accuracy of 98% [4]. Building on this, Sfayyih et al. conducted a literature review highlighting CNNs' superior performance over conventional machine learning models in respiratory classification tasks. Their review emphasized the effectiveness of pre-trained models such as ResNet and VGG, and explored architectural improvements like integrating Mixture of Experts (MoE) and LSTM for better accuracy and computational efficiency [17]. However, they also noted challenges in adversarial robustness, where small perturbations in input data can significantly mislead predictions—a concern also raised by Hinton et al. regarding neural network vulnerability [18].

Focusing on multi-channel input, Messner et al. introduced a CRNN-based model that achieved an F1-score of 92.4%, effectively modeling temporal and spatial characteristics of lung sounds [19]. Similarly, Haider et al. explored machine learning classifiers like SVM, LR, and kNN to distinguish COPD and healthy cases. While sound features alone yielded moderate performance (SVM: 83.6%), integrating spirometry data led to perfect classification (100% accuracy and AUC = 1), showcasing the potential of multimodal fusion [20][21]. The evaluation showed that classifiers using only lung sound features had moderate success—SVM reached 83.6% accuracy—but performance improved significantly when spirometry data was added. In this combined approach, both SVM and LR achieved 100% classification accuracy and an AUC of 1, outperforming all other methods.

III. Methodology

A. Data Collection

This study uses the Respiratory Sound Database from Kaggle, consisting of 5,539 WAV recordings from 126 patients across eight respiratory conditions: COPD, URTI, Bronchiectasis, Bronchiolitis, Pneumonia, LRTI, Asthma, and Healthy [22]. The recordings, collected in clinical settings using four devices (Littmann, AKGC417L, Meditron, and WelchAllyn), include metadata like age, gender, diagnosis, and recording location. Selected for its clinical authenticity, rich annotations, and size for training deep learning models, this dataset provides more reliable data than others that rely on simulated inputs. Its open academic license ensures ethical research use.

B. Exploratory Data Analysis

The data includes the following labels for respiratory conditions: COPD, LRTI, URTI, Pneumonia, Bronchiectasis, Bronchiolitis, Asthma, and Healthy.

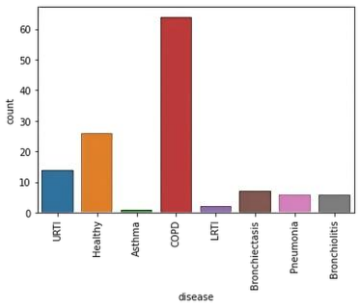

Fig. 1. Patient distribution

Fig. 1 shows the imbalance of the dataset used, whereas COPD is more dominant compared to the other patients. The numbers represent the total of patients rather than the total for each label.

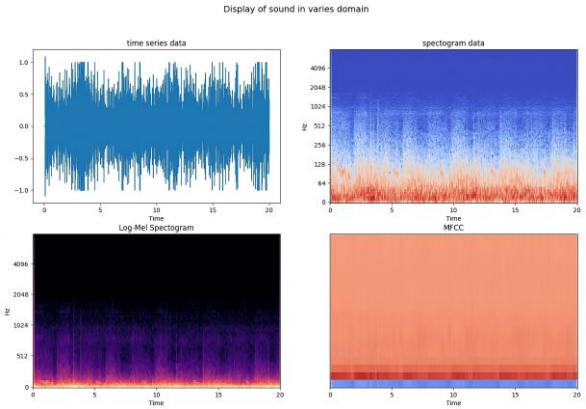

Fig. 2. Visualization of a COPD-diagnosed audio sample using Spectrograms and MFCCs.

Fig. 2 illustrates various audio representations derived from one of the respiratory sound samples in the dataset. The time series (top left) shows raw amplitude variations over time but lacks frequency insight. The spectogram (top right) reveals time-frequency information, with color indicating energy levels. The Log-Mel Spectogram (bottom left) applies the Mel scale and logarithmic amplitude, aligning with human hearing perception. Finally, MFCCs (bottom right) offer compact features that represent key timbral and perceptual characteristics of the sound.

C. Data Preprocessing

To start the preprocessing, audio files were merged to get the whole dataset. After merging, noise reduction techniques will be applied to minimize background noise. Next, normalization will standardize the recordings for consistency.

1) Augmentation

To address class imbalance, data augmentation techniques were applied using Gaussian noise addition and time stretching. Noise was added at random levels (0.005–0.015) to mimic real-world conditions without distorting respiratory patterns, while time stretching (rate 0.8–1.2) varied audio speed without affecting pitch. The stretched signals were normalized before inclusion in the augmented dataset. These steps were performed on the original, unaugmented data.

TABLE I. DATA BEFORE AUGMENTATION

Labels	Total
COPD	793
Non-COPD	127
Total	920

TABLE II. DATA AFTER AUGMENTATION

Labels	Total
COPD	793
Non-COPD	508
Total	1301

As shown in TABLE I, the original dataset had significantly fewer non-COPD audio samples compared to COPD. After applying noise addition and time stretching to the non-COPD class, TABLE II shows a more balanced class distribution. The dataset was then split 80:20, with 80% used for training and 20% for testing.

2) Noise Reduction

As illustrated in Fig. 2, the recordings are predominantly influenced by low-frequency noise. Pre-emphasis is a signal processing technique that amplifies high frequencies to balance the speech spectrum and improve feature extraction [23]. This method was applied to amplify abnormal respiratory sounds such as wheezing and crackles while suppressing the dominance of low-frequency interference.

3) Normalization

To ensure consistent amplitude across all audio recordings, normalization was applied. This process scales the audio signal so that its maximum absolute value is 1, resulting in amplitudes constrained within the range of -1 to 1 without altering the original waveform shape. Normalization is critical for stabilizing feature extraction, preventing bias caused by varying recording volumes, and improving the numerical stability of subsequent processing steps.

4) Padding

All audio samples are standardized to 5 seconds by padding shorter clips with silence and truncating longer ones. As a result, consistency is maintained in a feature extraction and model training.

D. Feature Extraction

The feature extraction phase will leverage advanced techniques to derive meaningful attributes from the audio recordings.

TABLE III. FEATURES USED IN THE STUDY

Features	Total of Coefficients
MFCC	13
Chromagram	12
Root Mean Square (RMS)	1
Spectral Centroid (SC)	1
Bandwidth	1
Spectral Roll-Off (SR)	1
Tonnetz (T)	6
Polynomial Coefficients (Poly)	1
Total	37

Features used are shown in TABLE III. Each feature contributes unique insights into the acoustic properties of lung sounds, as described below.

1) MFCC

Mel-Frequency Cepstral Coefficients (MFCCs) are employed in this study due to their proven effectiveness in capturing the perceptual and spectral characteristics of audio signals [24]. Recent research has further optimized MFCC parameters for respiratory disease detection, demonstrating their continued relevance in this domain [25]. As respiratory sounds often contain critical low-frequency components and subtle spectral patterns, MFCCs are particularly suitable for representing these features in a compact and noise-robust form.

2) Chromagram

Chromagram features creates audio respresentation by emphasizing tonal structures. Conditions like asthma, COPD, and airway obstructions generate characteristic sounds, such as wheezing and stridor, which have periodic energy in specific frequency regions. By mapping the spectrum into 12 pitch classes, Chromagram highlights

these tonal patterns while minimizing non-tonal noise and minor frequency variations [24][26].

3) Root Mean Square (RMS)

Root Mean Square (RMS) is a measure of the average energy or power of a signal over time, commonly used in audio processing to assess sound strength or loudness [26]. It helps in detecting weak breathing sounds, such as those from pleural effusion or shallow respiration, as well as sudden energy spikes linked to abnormal events like crackles, coughing, or difficulty breathing.

4) Spectral Centroid (SC)

The spectral centroid represents the "center of mass" of the spectrum, indicating where the bulk of the signal's frequency energy is concentrated [24]. The spectral centroid identifies the dominant frequency region, offering insights into whether energy is concentrated in lower or higher bands, which is critical for detecting abnormalities.

5) Bandwidth

Spectral bandwidth measures the spread of frequencies around the spectral centroid, providing an indication of how wide or narrow the frequency content of an audio signal is [24]. A signal with a narrow bandwidth is concentrated around a few frequencies, while a broader bandwidth suggests a more distributed frequency energy. In respiratory sound analysis, spectral bandwidth becomes particularly valuable because pathological sounds such as crackles or coarse breathing often exhibit a broader spread of energy across the frequency spectrum compared to normal, smooth breathing sounds.

6) Spectral Roll-Off (SR)

Spectral roll-off provides a practical way to describe how energy is distributed across the frequency spectrum of an audio signal by marking the frequency below which the bulk of the energy is concentrated [24]. In respiratory sound analysis, this becomes particularly important because abnormal breathing events, such as wheezing, tend to shift the energy toward higher frequencies. By measuring this shift, spectral roll-off helps distinguish between normal and pathological respiratory patterns.

7) Tonnetz (T)

Tonnetz is a feature representing harmonic relationships between pitch classes in a six-dimensional space, based on musical intervals like major and minor thirds [27]. It is valuable for respiratory audio, especially in analyzing tonal sounds like wheezing and stridor, which resemble musical tones. While traditional features focus on energy and frequency, Tonnetz highlights pitch relationships, which identifies tonal breathing sounds linked to lung diseases.

8) Zero-Crossing Rate (ZCR)

Zero-Crossing Rate (ZCR) is a time-domain feature that measures how frequently an audio signal crosses the zero-amplitude axis [24] . Zero-Crossing Rate (ZCR) provides a simple yet effective measure of signal frequency characteristics by quantifying oscillatory behavior. In respiratory audio, abnormal sounds such as crackles and wheezes typically exhibit higher zero-crossing activity compared to normal breathing.

9) Polynomial Coefficient (Poly)

Polynomial coefficients are calculated by fitting a simple curve to the overall shape of an audio signal's frequency spectrum [4]. This helps capture general patterns in how sound energy is distributed, without focusing on small details. In respiratory audio, these features are useful for detecting unusual patterns caused by conditions like wheezing or crackles. They are also less sensitive to noise or changes in recording, making them a reliable addition to other features such as MFCCs for identifying abnormal breathing.

E. Data Modelling

The modeling phase of this research will be developing a model for classifying samples into "COPD" and "non-COPD" categories, employing deep learning algorithms such as CNN + RNN, CNN + LSTM, and CNN + SVM.

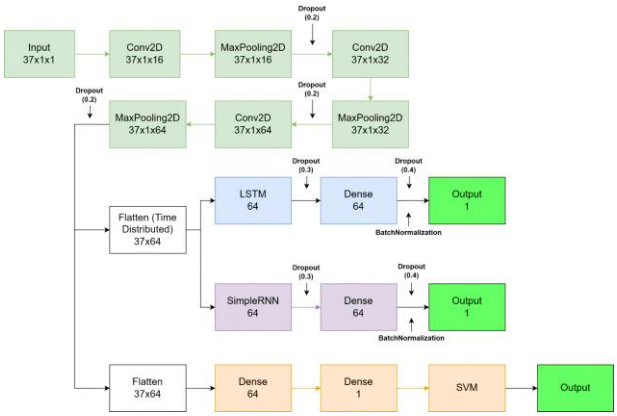

Fig. 3. The hybrid model architecture

Fig. 3 illustrates the architecture visualization. Seven models were implemented: three hybrid (CNN+SVM, CNN+RNN, CNN+LSTM) and four standalone (CNN, RNN, LSTM, SVM). To ensure fair evaluation, all models were trained with default parameters from TensorFlow/Keras and Scikit-learn, using 300 epochs, a batch size of 64, and dropout rates of 0.2 to 0.4 to prevent overfitting. The Adam optimizer was used, and early stopping was excluded to maintain consistent training duration. This standardized approach isolates architectural differences, minimizing performance bias from manual tuning.

F. Model Evaluation

Metrics including accuracy, precision, recall, specificity, F1-score, and ROC curves will be used to assess the models' performance. A comparative study will be carried out to determine which model performs the best.

IV. Result and Discussion

This section compares the models' performance based on key criteria: accuracy, precision, recall, and F1-score, aiming to highlight their advantages and disadvantages in different scenarios.

TABLE IV. PERFORMANCE EVALUATION

Evaluation		CNN+LSTM	CNN+SVM	CNN+RNN	CNN	LSTM	SVM	RNN
Precision	COPD	0.90	0.87	0.85	0.93	1.00	0.92	0.95
	Non-COPD	0.89	0.78	0.52	0.84	0.89	0.85	0.55
Recall	COPD	0.94	0.86	0.49	0.89	0.93	0.90	0.51
	Non-COPD	0.84	0.79	0.86	0.90	1.00	0.88	0.96
F1-Score	COPD	0.92	0.86	0.62	0.91	0.96	0.91	0.66
	Non-COPD	0.87	0.78	0.65	0.87	0.94	0.86	0.70

As shown in TABLE IV, CNN+LSTM achieved the best overall performance among hybrid models, combining CNN's spatial feature extraction with LSTM's ability to capture long-term temporal patterns. However, the improvement over base models like LSTM and CNN was minimal, indicating limited benefits of hybridization on clean data. CNN+SVM also performed well, leveraging CNN's representation learning and SVM's decision-boundary precision, though similar to base models. CNN+RNN performed the worst, especially in COPD recall, likely due to RNN's struggle with maintaining relevant temporal information over long sequences. These results suggest that CNN-based hybrids, particularly with LSTM, can be useful but should be applied selectively.

TABLE V. STATISTICAL TEST

Statistic Test		CNN+LSTM	CNN+SVM	SNN+RNN	CNN	LSTM	SVM	RNN
Accuracy	Mean	0.900	0.832	0.635	0.898	0.954	0.893	0.681
	Std	0.019	0.023	0.029	0.019	0.013	0.019	0.029
	95% CI	0.862-0.935	0.782-0.877	0.575-0.694	0.858-0.931	0.927-0.977	0.858-0.931	0.625-0.743
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Precision	Mean	0.893	0.777	0.515	0.843	0.892	0.848	0.549
	Std	0.032	0.040	0.039	0.035	0.030	0.035	0.038
	95% CI	0.765-0.911	0.693-0.854	0.426-0.590	0.771-0.908	0.833-0.949	0.774-0.913	0.472-0.623
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Recall	Mean	0.842	0.792	0.861	0.901	1.000	0.881	0.960
	Std	0.038	0.040	0.035	0.031	0.000	0.032	0.020
	95% CI	0.765-0.911	0.716-0.867	0.792-0.928	0.839-0.958	1.000-1.000	0.812-0.938	0.917-0.991
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ROC AUC	Mean	0.972	0.920	0.735	0.971	0.993	0.968	0.821
	Std	0.009	0.016	0.029	0.008	0.004	0.009	0.024
	95% CI	0.953-0.987	0.887-0.950	0.680-0.792	0.953-0.985	0.984-0.999	0.950-0.983	0.774-0.865
	P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000

TABLE V summarizes model performance based on accuracy, precision, recall, ROC AUC, and statistical metrics like standard deviation, 95% confidence intervals, and p-values. The CNN+LSTM hybrid model achieved the best results, showing the advantage of combining spatial and temporal features. However, it only slightly outperformed the simpler LSTM model, suggesting limited benefit relative to added complexity. CNN+SVM also performed well but was comparable to its individual components. CNN+RNN had the weakest performance, likely due to RNN's limitations with long-term dependencies. Narrow confidence intervals in top-performing models indicate consistent and reliable results. Wider intervals in weaker models like CNN+RNN reflect less stable performance. Additionally, all p-values were below 0.001, confirming that performance differences between models are statistically significant and not due to random variation. In summary, while CNN-based hybrids—particularly with LSTM—can be effective, strong base models may already provide sufficient performance depending on dataset complexity.

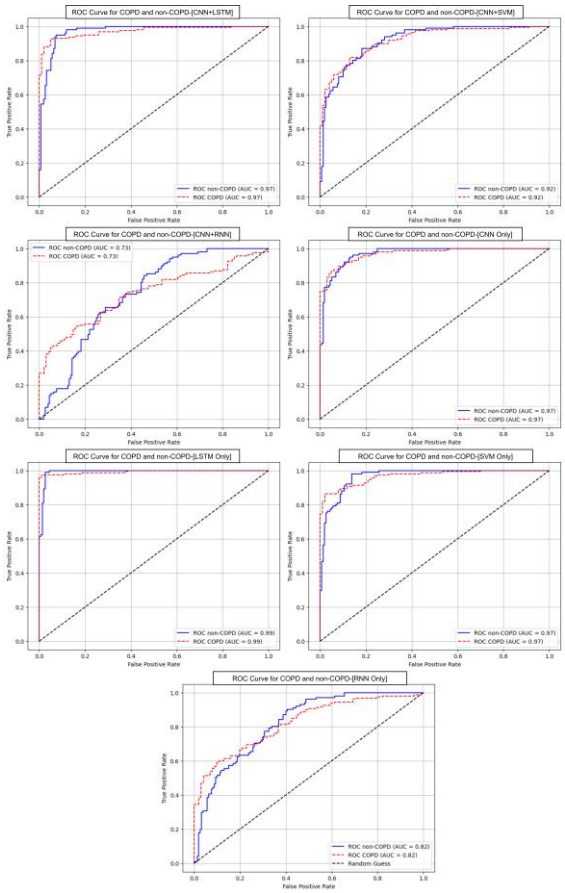

Fig. 4. ROC Curve of all the models

Fig. 4 presents the ROC curves for each model in distinguishing between COPD and non-COPD cases. The CNN+LSTM hybrid achieved excellent performance, indicating high sensitivity and specificity across thresholds. However, the standalone LSTM model slightly outperformed it, and CNN alone matched CNN+LSTM exactly, reinforcing that both base models are already highly effective. SVM with all features also performed well, though it remains slightly below the deep learning models. In contrast, the CNN+RNN and RNN-only models demonstrated notably lower AUC values, with ROC curves that deviate from the ideal top-left corner, reflecting weaker and less consistent classification. These results visually support the statistical findings, showing that while CNN+LSTM is a strong hybrid, its benefit is marginal when base models like CNN or LSTM already perform near optimally.

V. CONCLUSION

The CNN-LSTM model achieved the highest overall performance across all metrics—precision, recall, F1-score, and accuracy—with stable convergence, a recall of 0.91 for COPD, and peak accuracy at 0.88. This suggests its strong capability in capturing temporal dependencies within respiratory sounds, making it the most reliable architecture for COPD classification. CNN-SVM also showed consistent

performance despite slower convergence, while CNN-RNN exhibited instability and signs of overfitting. ROC analysis supported these findings, confirming CNN-LSTM as the top-performing model, followed by CNN-SVM. Importantly, all models—particularly CNN-LSTM—maintained robust performance despite the presence of four different recording devices in the dataset, indicating a promising level of resilience to device variability, though more targeted generalization.

Despite its strengths, CNN-LSTM is computationally intensive and less suited for deployment on edge or mobile devices. Future work should explore model compression or lightweight architectures to enable real-time applications in low-resource environments. Additionally, the development of privacy-preserving frameworks such as federated learning or encryption is essential for handling sensitive medical data securely. to improve fairness and generalizability, future work should balance the dataset, diversify recording conditions, and reduce potential bias through more inclusive training.

AUTHOR CONTRIBUTION

C.C.R. conducted the analysis, interpreted the results, and drafted the manuscript. C.M.L. conceptualized the study, designed the framework, and conducted the experiments. Supervised by A.A.S.G. and J.J.T., who provided guidance and feedback. All authors approved the final manuscript.

AVAILABILITY DATA AND MATERIALS

This study used the "Respiratory Sound Database" by the School of Health Sciences at the University of Aveiro (ESSUA) and the Aristotle University of Thessaloniki (AUTH) [22].

ACKNOWLEDGEMENT

We would like to express our gratitude to the GeoEco-AI research interest group for providing valuable research ideas and computational resources. We also acknowledge the support of the Bina Nusantara University International Research Fund (PIB) under Proposal No. No: 081/VRRTT/IV/2025, titled "Medical Image Analysis with Deep Learning: Enhancing Precision in Health Diagnosis."

REFERENCES

[1] J. R. Hurst et al., "Prognostic risk factors for moderate-to-severe exacerbations in patients with chronic obstructive pulmonary disease: A systematic literature review," Respiratory Research, vol. 23, no. 1, Aug. 2022. doi:10.1186/s12931-022-02123-5

[2] C.-Z. Chen et al., "Life expectancy (le) and loss-of-le for patients with chronic obstructive pulmonary disease," Respiratory Medicine, vol. 172, p. 106132, Oct. 2020. doi:10.1016/j.rmed.2020.106132

[3] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," EURASIP Journal on Image and Video Processing, 2017.

[4] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "A Neural Network-Based Method for Respiratory Sound Analysis and Lung Disease Detection," Applied Sciences, 2022.

[5] M. Milicevic, I. Mazic, and M. Bonkovic, "Classification Accuracy Comparison of Asthmatic Wheezing Sounds," Computational Science and Systems Engineering.

[6] S.-Y. Jung, C.-H. Liao, Y.-S. Wu, S.-M. Yuan, and C.-T. Sun, "Efficiently classifying lung sounds through depthwise separable CNN models with fused STFT and MFCC features," Diagnostics, vol. 11, no. 4, p. 732, Apr. 2021. doi:10.3390/diagnostics11040732

[7] M. Milicevic, I. Mazic, and M. Bonkovic, "Asthmatic Wheezes Detection - What Contributes the Most to the Role of MFCC in Classifiers Accuracy?"

[8] C. F. Vogelmeier et al., "Goals of COPD treatment: Focus on symptoms and exacerbations," Respiratory Medicine, vol. 166, p. 105938, May 2020. doi:10.1016/j.rmed.2020.105938

[9] C. F. Vogelmeier et al., "Global strategy for the diagnosis, management, and Prevention of Chronic Obstructive Lung Disease 2017 report. Gold Executive Summary," American Journal of Respiratory and Critical Care Medicine, vol. 195, no. 5, pp. 557–582, Mar. 2017. doi:10.1164/rccm.201701-0218pp

[10] Q. Kong et al., "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880–2894, 2020.

[11] E. U. R. Mohammed, N. R. Soora, and S. W. Mohammed, "A Comprehensive Literature Review on Convolutional Neural Networks," *University of Windsor Computer Science Publications*, 2022. https://scholar.uwindsor.ca/computersciencepub/58.

[12] J. Breebaart and M. McKinney, "Features for Audio Classification," in Audio Signal Processing for Next-Generation Multimedia Communication Systems, 2004, pp. 1-15. https://doi.org/10.1007/978-94-017-0703-9.

[13] S. Hershey et al., "CNN architectures for large-scale audio classification," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131–135, Mar. 2017. doi:10.1109/icassp.2017.7952132

[14] M. Massoudi, S. Verma, and R. Jain, "Urban sound classification using CNN," in Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT), 2021, pp. 583–588. doi: 10.1109/ICICT50816.2021.9358621.

[15] A. Ashar, M. S. Bhatti, and U. Mushtaq, "Speaker identification using a hybrid CNN-MFCC approach," in Proc. 2020 Int. Conf. Emerging Trends in Smart Technol. (ICETST), 2020, pp. 1–6. doi: 10.1109/ICETST49965.2020.9080730.

[16] Srivastava, et al., "Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease," PeerJ Comput. Sci., vol. 7, 2021. doi: 10.7717/peerj-cs.369

[17] A.H. Sfayyih, et al., "Acoustic-based deep learning architectures for lung disease diagnosis: A comprehensive overview," Diagnostics, vol. 13, no. 10, p. 1748, 2023. doi: 10.3390/diagnostics13101748

[18] G.E. Hinton, et al., "Improving neural networks by preventing co-adaptation of feature detectors," 2012. https://doi.org/10.48550/arXiv.1207.0580

[19] E. Messner et al., "Multi-channel lung sound classification with convolutional recurrent Neural Networks," Computers in Biology and Medicine, vol. 122, p. 103831, Jul. 2020. doi:10.1016/j.compbiomed.2020.103831

[20] N. S. Haider et al., "Respiratory sound based classification of chronic obstructive pulmonary disease: A risk stratification approach in machine learning paradigm," Journal of Medical Systems, vol. 43, no. 8, Jun. 2019. doi:10.1007/s10916-019-1388-0

[21] Mineshita, M. et al., The correlation between lung sound distribution and pulmonary function in COPD patients. PLoS ONE 9(9):e107506, 2014.

[22] B. M. Rocha *et al.*, "An open access database for the evaluation of Respiratory Sound Classification algorithms," *Physiological Measurement*, vol. 40, no. 3, p. 035001, Mar. 2019. doi:10.1088/1361-6579/ab03ea

[23] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," *Proc. Canadian Conf. Electrical and Computer Engineering*, pp. 1062–1065, 1995. doi: 10.1109/CCECE.1995.528521.

[24] C. Bisogni, V. Loia, M. Nappi, and C. Pero, "Acoustic features analysis for explainable machine learning-based audio spoofing detection," Comput. Vis. Image Underst., vol. 249, 104145, 2024. doi: 10.1016/j.cviu.2024.104145.

[25] Yan, Y., Simons, S. O., van Bemmel, L., Reinders, L., Franssen, F. M. E., & Urovi, V. (2024). Optimizing MFCC parameters for the automatic detection of respiratory diseases. Applied Acoustics, 228, 110299. https://doi.org/10.1016/j.apacoust.2024.110299

[26] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," Appl. Acoust., vol. 158, 107020, 2020. doi: 10.1016/j.apacoust.2019.107020.

[27] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conf. (SciPy 2015)*, S. Benthall and S. Rostrup, Eds., 2015, pp. 18–25. [Online]. Available: https://doi.org/10.25080/Majora-7b98e3ed-003