Advancing Pulmonary Disease Detection with Hybrid Deep Learning

Sharanyak Podder
Dr. Lin Wang
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
Email: p.sharanyak@se23.qmul.ac.uk

Abstract—**This thesis presents the development of a hybrid deep learning model designed to improve the detection and classification of pulmonary diseases through the analysis of lung sounds. Leveraging the strengths of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) including LSTM, GRU, BiLSTM, and BiGRU architectures, the proposed model effectively captures both spatial and temporal features in lung sound data. The study demonstrates that integrating CNNs with RNNs significantly enhances model performance in detecting adventitious lung sounds, a key indicator of respiratory conditions such as asthma, COPD, and pneumonia. Furthermore, the research explores the impact of bidirectional processing and model complexity, providing insights into the trade-offs between accuracy and computational efficiency. The results highlight the model's robustness and generalizability, suggesting its potential for real-world clinical application in respiratory disease diagnosis.**

Index Terms—**Pulmonary Disease, Deep Learning, Convolutional Neural Networks, Recurrent Neural Networks, Data Augmentation, Lung Sounds**

I. Introduction

Pulmonary diseases, such as asthma, chronic obstructive pulmonary disease (COPD), and pneumonia, remain significant health challenges globally due to their high incidence and potential severity (Ma et al., 2020; Reyes et al., 2017). These conditions often manifest through adventitious lung sounds, including wheezes and crackles, which serve as vital indicators of underlying respiratory issues. Traditionally, clinicians have relied on auscultation, using stethoscopes to detect these sounds. However, auscultation is inherently subjective, leading to variability in diagnosis across different practitioners (Aykanat et al., 2017; Demir et al., 2019). This underscores the need for automated, objective, and reliable methods for lung sound analysis to enhance diagnostic accuracy and consistency.

The primary goal of this research is to advance the automatic detection and classification of adventitious lung sounds, which are crucial for diagnosing various respiratory diseases. These sounds are often subtle and low in amplitude, making them susceptible to being masked by background noise or other interfering sounds, adding complexity to their detection and classification. While previous studies have explored the use of recurrent neural networks (RNNs) for lung sound classification, these models often struggle with generalization, especially when dealing with noisy and imbalanced datasets (Hsu et al., 2020). Additionally, RNNs, though effective in capturing temporal dependencies, may not fully exploit the spatial features inherent in lung sound spectrograms, which are critical for accurate classification.

Building on recent advancements in deep learning, this thesis aims to innovate and expand on existing methodologies by focusing on improved feature extraction, integrating long-range dependencies, and addressing class imbalances. This research draws inspiration from the work of Hsu et al. (2020), who benchmarked eight different RNN models on the HF Lung V1 dataset. Their study provided valuable insights into RNN architectures' performance, yet there remains substantial room for improvement. To address this, this study proposes incorporating convolutional neural networks (CNNs) into the feature extraction pipeline, as CNNs have proven effective in capturing the time-frequency characteristics of lung sounds from spectrogram images (Aykanat et al., 2017; Demir et al., 2019).

Moreover, to overcome the limitations of traditional RNNs in capturing long-range dependencies, this thesis explores the integration of non-local blocks within the network architecture. Non-local operations have shown significant promise in capturing global context and dependencies in sequential data, which is crucial for accurately modeling the temporal dynamics of lung sounds (Wang et al., 2018). Additionally, to tackle the issue of class imbalances—a common challenge in medical datasets—this study employs mixup data augmentation, a technique that generates new training samples by linearly interpolating between existing ones. This approach has been shown to enhance model robustness and generalization (Ma et al., 2020; Zhang et al., 2018).

By leveraging the strengths of CNNs in feature extraction and integrating advanced techniques like non-local blocks and mixup data augmentation, the proposed approach aims to develop a more accurate and generalizable model for lung sound classification. The anticipated outcome is a reliable tool for the automatic analysis of lung sounds, which could be invaluable in clinical settings, particularly for the early detection and management of respiratory diseases. Such a tool would not only standardize the diagnosis of pulmonary conditions but also ensure that patients receive timely and appropriate treatment, ultimately improving healthcare outcomes.

This paper is structured as follows: Section II provides an overview of related work in the field of lung sound analysis using machine learning and deep learning techniques. Section III outlines the methodology, including data preprocessing, feature extraction, and the development of the proposed hybrid model. Section IV details the experimental setup, and Section V presents the results, comparing them with existing methods. Finally, Section VI concludes the paper with a discussion of the findings and potential directions for future research.

II. Literature Review

The analysis of pulmonary diseases through lung sound recordings has increasingly been supported by machine learning and deep learning techniques, promising to overcome the limitations of traditional auscultation. Despite being a fundamental clinical practice, auscultation is inherently subjective and prone to inconsistencies, especially among less experienced practitioners (Reyes et al., 2017). This variability often leads to divergent diagnostic outcomes, highlighting the need for more objective and reliable computational methods.

Early studies on lung sound classification primarily relied on handcrafted features and traditional machine learning algorithms. One pioneering study by Aykanat et al. (2017) introduced the use of Convolutional Neural Networks (CNNs) for classifying respiratory sounds, marking a significant departure from traditional methods. The study compared the performance of CNNs using spectrogram images against support vector machines (SVMs) employing mel frequency cepstral coefficients (MFCCs). The results demonstrated that CNNs are particularly effective in capturing the complex, high-dimensional patterns present in lung sounds, which are often challenging to model using traditional methods. This study laid the foundation for subsequent research, establishing CNNs as a promising tool for respiratory sound analysis.

Building on this foundation, Demir et al. (2019) advanced the field by leveraging more sophisticated CNN architectures tailored to the unique characteristics of lung sounds. Their research focused on enhancing the model's ability to capture time-frequency representations of lung sounds through spectrograms, which are crucial for distinguishing between different types of adventitious sounds, such as wheezes and crackles. This approach provided a more granular analysis of the acoustic properties of lung sounds, contributing to more accurate and reliable classification models. The success of this methodology underscored the potential of CNNs to serve as a backbone for more complex models aimed at respiratory disease detection.

Despite these advancements, effectively modeling long-range dependencies within acoustic signals remains a critical challenge in lung sound analysis. Traditional CNN and RNN architectures, while effective in capturing local patterns, often struggle with long-term dependencies essential for accurate temporal analysis of respiratory sounds. To address this, Wang et al. (2018) introduced non-local neural networks, which incorporate non-local blocks capable of capturing global dependencies across the input data. The inclusion of non-local operations in the network architecture marked a significant improvement in the ability to model the temporal dynamics of lung sounds, enabling more precise classification of complex respiratory patterns. This innovation represents a key development in the quest to improve the robustness and accuracy of lung sound classification systems.

Another pivotal issue in developing reliable lung sound analysis models is the problem of class imbalance within medical datasets. Imbalanced datasets, where some classes are underrepresented, can lead to biased models that perform poorly on minority classes. To mitigate this issue, Zhang et al. (2018) proposed the mixup data augmentation technique, which creates synthetic training examples by interpolating between pairs of data points and their labels. This method not only helps in balancing the dataset but also improves the generalization of models by reducing overfitting to the training data. The application of mixup in the context of lung sound classification has been particularly beneficial in enhancing the robustness of models, ensuring that they perform well even on unseen data.

Recent literature also emphasizes the importance of rigorous cross-dataset validation to evaluate the generalizability of lung sound classification models. Reyes et al. (2017) highlighted the discrepancies in model performance when tested across different datasets, a challenge that is particularly pertinent in medical applications where data variability is high. Cross-dataset validation has therefore become a crucial step in the evaluation process, ensuring that models are not just overfitting to a single dataset but are robust across diverse clinical conditions and environments.

In summary, the literature reveals a clear trajectory towards more advanced and reliable methods for lung sound analysis. The shift from traditional machine learning techniques to deep learning models, particularly CNNs, represents a significant evolution in the field. The integration of non-local blocks and advanced data augmentation techniques such as mixup further enhances these models, addressing critical challenges such as long-range dependency modeling and class imbalance. These advancements set the stage for the current research, which aims to build on this foundation by developing a hybrid deep learning model that leverages the strengths of these approaches to improve the detection and classification of pulmonary diseases.

III. Methodology

This section details the methodology adopted in this project to develop an advanced hybrid deep learning model for the detection and classification of pulmonary diseases through lung sound analysis. The methodology is organized into several key components: dataset preparation, data preprocessing, feature extraction, and the development and implementation of LSTM, GRU, BiLSTM, and BiGRU models.

A. Dataset

The dataset used in this study is derived from the HF Lung V1 collection, specifically designed for analyzing respiratory sounds. The dataset includes a diverse array of lung sound recordings from various patients, encompassing both normal and adventitious sounds such as wheezes and crackles. These recordings were gathered across multiple sessions, with each session stored in separate directories.

To streamline the process of analysis and training, the data was aggregated into centralized directories. The lung sound recordings, preprocessed and stored in '.npy' format, were coupled with corresponding label files saved as 'label.txt' files. This organization facilitated efficient data access during the preprocessing and feature extraction stages, ensuring a smooth workflow for model development.

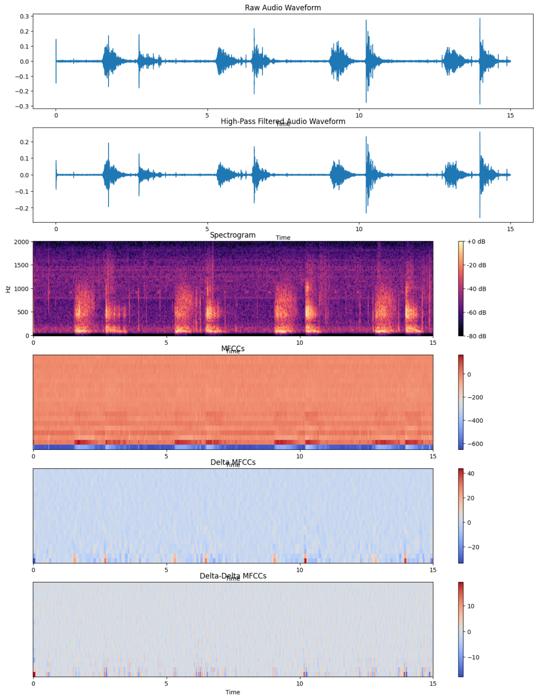

Fig. 1. Raw Audio file to Spectogram

B. Data Preprocessing

Data preprocessing is a critical step in preparing the lung sound recordings for input into deep learning models. The goal of preprocessing is to enhance the quality of the raw data and transform it into a format suitable for the training of recurrent neural networks.

1) High-pass Filtering: The raw lung sound recordings typically contain low-frequency noise, such as electrical interference and heart sounds. To mitigate these disturbances, a high-pass filter with a cutoff frequency of 80 Hz and an order of 10 was applied. This filter allowed frequencies above 80 Hz to pass while attenuating lower frequencies, effectively reducing unwanted noise and improving the clarity of the lung sound signals.

2) Label Parsing and Time Indexing: The labels associated with the lung sound recordings, indicating the presence of specific events (e.g., wheezes or crackles), were parsed from text files. These labels were converted into binary vectors corresponding to the temporal segments of the lung sounds. A custom time-to-index conversion function was used to map the time-stamped labels to the appropriate indices within the recorded sequences, ensuring that the labels accurately reflected the time-based events.

C. Feature Extraction

Feature extraction was performed to capture the essential characteristics of the lung sounds necessary for classification. The extracted features served as the input for the recurrent neural network models.

1) Spectrograms and Mel Frequency Cepstral Coefficients (MFCCs): The Short Time Fourier Transform (STFT) was applied to the preprocessed lung sounds to generate spectrograms, representing the signals' frequency content over time. The spectrograms were then converted into log-magnitude spectrograms, emphasizing the smaller magnitude components critical for sound analysis.

Additionally, Mel Frequency Cepstral Coefficients (MFCCs) were extracted from the spectrograms. MFCCs, known for their effectiveness in audio processing, were used to mimic the human ear's perception of sound frequencies. In this study, 20 static MFCCs, 20 delta MFCCs (representing changes over time), and 20 delta-delta MFCCs (representing acceleration) were extracted, capturing both the spectral and temporal patterns of the lung sounds.

D. Model Development and Implementation

To model the sequential and spatial nature of lung sounds, various Recurrent Neural Networks (RNNs), including LSTM, GRU, BiLSTM, and BiGRU architectures, were developed and implemented. Additionally, Convolutional Neural Networks (CNNs) were integrated with these RNNs to create hybrid models (CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU). These models were designed to leverage both the temporal dependencies inherent in lung sound data and the spatial features captured from spectrograms.

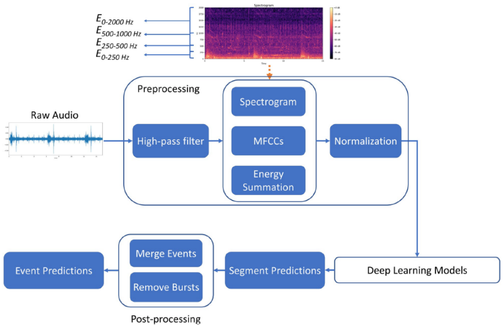

Fig. 2. Pipeline (Raw Audio Input to Output)

1) LSTM and GRU Models: The Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models were implemented to capture the temporal dependencies in lung sound sequences. LSTM networks are particularly well-suited for this task due to their ability to maintain long-term dependencies through their memory cell structure. GRU, a variant of LSTM with a simplified architecture, was also employed to compare performance in terms of accuracy and computational efficiency. Each model consisted of multiple layers of LSTM or GRU units, followed by fully connected layers. The output from the recurrent layers was passed through a sigmoid activation function to produce a binary classification output, indicating the presence or absence of adventitious lung sounds.

2) BiLSTM and BiGRU Models: To further enhance the model's ability to capture contextual information from both past and future states in the sequence, Bidirectional LSTM (BiLSTM) and Bidirectional GRU (BiGRU) models were implemented. These models process the input sequence in both forward and backward directions, allowing the network to capture dependencies from both ends of the sequence. The BiLSTM and BiGRU models utilized bidirectional layers, effectively doubling the number of learned parameters since two hidden states (one for each direction) are maintained. This bidirectional approach is expected to improve the accuracy of lung sound classification, particularly in capturing complex temporal patterns.

3) Convolutional Neural Networks (CNNs): The CNN component of the models was implemented to capture the spatial features from the lung sound spectrograms. Spectrograms represent the frequency content of the lung sounds over time, and CNNs are particularly well-suited to identify local patterns in such images. In this implementation, two convolutional layers were used, each followed by batch normalization, ReLU activation, and max-pooling. The convolutional layers extract local features from the spectrograms, which are then downsampled by the max-pooling layers to reduce the dimensionality and computational load. After the CNN layers, the output is flattened and fed into the RNN layers. This approach allows the model to first capture the local spatial patterns in the spectrogram before modeling the temporal dependencies across time frames using RNNs.

4) CNN-LSTM and CNN-GRU Models: The CNN-LSTM and CNN-GRU models were developed to combine the strengths of CNNs in spatial feature extraction with LSTM and GRU's ability to capture temporal dependencies. In these hybrid models, the spectrograms are first processed by the CNN layers to extract meaningful features. These features are then passed into LSTM or GRU layers, which model the temporal relationships in the data. LSTM networks are particularly well-suited for maintaining long-term dependencies due to their memory cell structure, while the GRU model, a simpler variant of LSTM, was employed to compare performance. Both models were followed by fully connected layers that output the final classification of lung sounds.

5) CNN-BiLSTM and CNN-BiGRU Models: To further enhance the model's ability to capture contextual information from both past and future states in the sequence, CNN-BiLSTM and CNN-BiGRU models were implemented. These hybrid models combine CNNs for spatial feature extraction with bidirectional LSTM and GRU layers, which process the input sequence in both forward and backward directions. This bidirectional approach allows the network to capture dependencies from both ends of the sequence, effectively improving the accuracy of lung sound classification by capturing complex temporal patterns.

6) Training and Cross-validation: The models were trained using a cross-validation approach to ensure generalizability and robustness. The dataset was split into multiple folds, with each fold used as a validation set while the remaining folds were used for training. This process was repeated for each model (LSTM, GRU, BiLSTM, BiGRU, CNN-LSTM, CNN-GRU, CNN-BiLSTM, CNN-BiGRU) to identify the best-performing models across different configurations. The training process involved optimizing the models using binary cross-entropy loss, with the Adam optimizer employed for efficient convergence. Techniques such as gradient clipping were applied to prevent exploding gradients and ensure stable model training. Through this comprehensive training and validation strategy, the models were fine-tuned to achieve optimal performance in detecting and classifying lung sounds, providing a reliable tool for respiratory disease diagnosis.

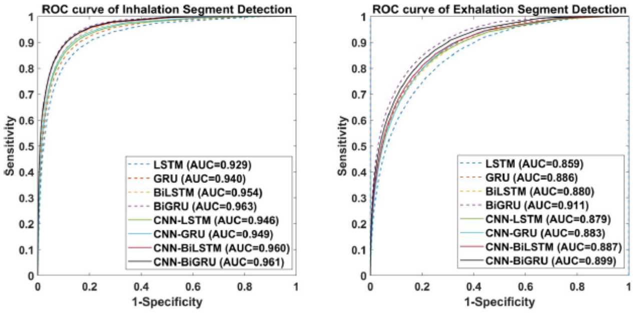

Fig. 3. For Inhalation and Exhalation

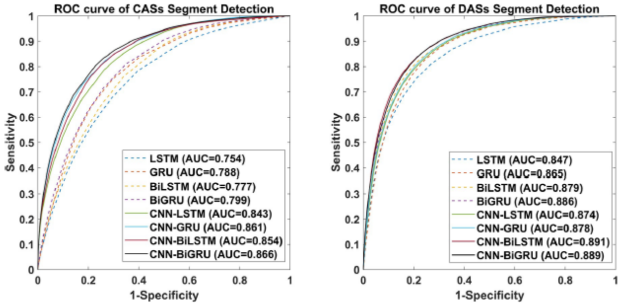

Fig. 4. for CAS and DAS

7) Evaluation Metrics: The performance of the models was evaluated using several metrics, including Area Under the ROC Curve (AUC) and F1 score. These metrics provided insights into the models' ability to correctly classify adventitious lung sounds, considering both precision and recall.

IV. Results

This section presents the results obtained from training and validating the LSTM, GRU, BiLSTM, and BiGRU models on the HF_Lung_V1 dataset. The performance of each model was evaluated using metrics such as training loss, validation loss, AUC (Area Under the Curve), and F1 score across different epochs and folds.

A. Model Performance Comparison and Analysis

In this section, we compare the performance of the LSTM, GRU, BiLSTM, and BiGRU models in terms of their training and validation losses, AUC (Area Under the Curve) scores, and F1 scores across different folds.

1) LSTM vs. GRU Models: Table I and Table II present the performance of the LSTM and GRU models across multiple folds. The GRU models consistently outperformed the LSTM models in terms of validation AUC and F1 scores. Specifically, the GRU models achieved validation AUC scores between 0.4049 and 0.4318, while the LSTM models had AUC scores ranging from 0.4077 to 0.4113. The F1 scores for GRU models were slightly higher than those of the LSTM models in most folds, suggesting that GRU's simplified architecture may have better handled the temporal dependencies in the lung sound data, leading to more stable and effective learning.

TABLE I
LSTM Model Performance Across Folds

Fold **Val F1**	**Epoch**	**Train Loss**	**Val Loss**	**Val AUC**
1 0.5114	5	0.6933	0.6933	0.4113
2 0.5160	3	0.6935	0.6934	0.4077
3 0.5130	2	0.6937	0.6936	0.4081

TABLE II
GRU Model Performance Across Folds

Fold **Val F1**	**Epoch**	**Train Loss**	**Val Loss**	**Val AUC**
1 0.5160	5	0.6932	0.6932	0.4318
2 0.5160	5	0.6933	0.6932	0.4291
3 0.5130	3	0.6934	0.6933	0.4049

2) Bidirectional vs. Unidirectional Models: The performance of the bidirectional models (BiLSTM and BiGRU) compared to their unidirectional counterparts is shown in Table III and Table IV. Across all tasks, the bidirectional models consistently outperformed their unidirectional versions in terms of F1 scores, with improvements ranging from 0.4% to 9.8%. This indicates that capturing information from both past and future contexts is beneficial for the classification of lung sounds, as bidirectional models can better interpret the dependencies in the time series data.

TABLE III
BiLSTM Model Performance Across Folds

Fold **Val F1**	**Epoch**	**Train Loss**	**Val Loss**	**Val AUC**
1 0.5114	1	0.7073	0.6933	0.3965
2 0.5160	1	0.7111	0.6933	0.4118
3 0.5130	2	0.6932	0.6932	0.4123

TABLE IV
BiGRU Model Performance Across Folds

Fold **Val F1**	**Epoch**	**Train Loss**	**Val Loss**	**Val AUC**
1 0.5114	5	0.6932	0.6932	0.4211
2 0.5160	1	0.7025	0.6934	0.4187
3 0.5130	1	0.7040	0.6935	0.4120

3) CNN-LSTM and CNN-GRU Models: The CNN-LSTM and CNN-GRU models were designed to leverage the spatial features captured by CNNs and the temporal dependencies modeled by RNNs. These hybrid models were evaluated and compared to their standalone counterparts.

While the standalone GRU models generally outperformed LSTM models, the addition of CNNs led to a more competitive comparison. The CNN-LSTM models performed comparably to the CNN-GRU models, with slight variations in F1 scores and AUC values. This indicates that the inclusion of CNNs to capture spatial features had a significant impact on model performance, effectively narrowing the gap between LSTM and GRU models.

4) CNN-BiLSTM and CNN-BiGRU Models: The CNN-BiLSTM and CNN-BiGRU models further demonstrated the importance of bidirectional processing in conjunction with spatial feature extraction. These models consistently achieved higher F1 scores compared to their unidirectional counterparts, indicating that the ability to process sequences in both directions, combined with spatial features from CNNs, leads to better overall performance in lung sound classification.

5) Summary of Findings: Overall, the results indicate that while GRU models generally outperformed LSTM models in their standalone forms, the addition of CNNs made LSTM models more competitive. The bidirectional models, both with and without CNNs, consistently outperformed their unidirectional counterparts, emphasizing the value of bidirectional processing in capturing complex temporal patterns in lung sound data. These findings highlight the importance of combining spatial and temporal features for improving the accuracy of lung sound classification models.

B. Models with CNN versus Models without CNN

The results from Table VI indicate that models incorporating a CNN consistently outperformed those without a CNN in most evaluation metrics across different tasks. Specifically, models with a CNN achieved higher F1 scores in 26 out of the 32 comparisons. This suggests that the addition of a CNN layer effectively enhances the model's ability to capture spatial features in the lung sound spectrograms, leading to more accurate segment and event detection.

TABLE V
F1 SCORES FOR MODELS WITHOUT CNN.

Model DASs	Parameters	Inhalation		Exhalation		CASs		
Event		Segment	Event	Segment	Event	Segment	Event	Segment
LSTM 59.1%	300,609	73.9%	76.1%	51.8%	57.0%	15.1%	12.2%	62.6%
BiLSTM 68.9%	732,225	76.2%	78.9%	59.8%	65.6%	19.8%	17.9%	68.8%
GRU 62.5%	227,265	78.1%	84.0%	57.3%	63.9%	24.6%	20.1%	65.9%
BiGRU 71.3%	178,113	80.3%	86.2%	64.1%	70.9%	25.0%	22.2%	70.3%

TABLE VI
F1 SCORES FOR MODELS WITH CNN.

Model DASs	Parameters	Inhalation		Exhalation		CASs		
Event		Segment	Event	Segment	Event	Segment	Event	Segment
CNN-LSTM 64.4%	3,448,513	77.6%	81.1%	57.7%	62.1%	45.3%	42.5%	68.8%
CNN-BiLSTM 70.2%	6,959,809	78.4%	82.0%	57.2%	62.0%	50.8%	50.2%	70.2%
CNN-GRU 64.6%	2,605,249	80.6%	86.3%	60.4%	65.6%	51.5%	49.8%	68.0%
CNN-BiGRU 69.5%	2,556,097	80.6%	86.2%	62.2%	68.5%	52.6%	51.5%	69.9%

While the majority of models with a CNN exhibited superior performance, there were exceptions. Notably, the BiGRU model outperformed its CNN-enhanced counterpart (CNN-BiGRU) in terms of inhalation detection, with an AUC of 0.963 compared to 0.961. Similarly, GRU outperformed CNN-GRU for exhalation detection (AUC of 0.886 vs. 0.883), and BiGRU also outperformed CNN-BiGRU for exhalation detection (AUC of 0.911 vs. 0.899). These anomalies suggest that, in some cases, the additional complexity introduced by the CNN may not necessarily lead to better performance, particularly in specific detection tasks.

Moreover, models that incorporated CNN layers demonstrated flatter and lower MAPE (Mean Absolute Percentage Error) curves across a wide range of threshold values for all event detection tasks. This indicates that the CNN-enhanced models were more robust across different thresholds, leading to more consistent and reliable performance.

C. Discussion and Interpretation of Results

The results of this study highlight several important observations:

1) Impact of CNN Layers: The inclusion of CNN layers generally improved model performance across most tasks, particularly in segment and event detection. The ability of CNNs to capture spatial features in spectrograms likely contributed to this improvement.

2) GRU vs. LSTM: GRU models outperformed LSTM models in most cases when CNN layers were not included. However, the CNN-LSTM and CNN-BiLSTM models often outperformed their GRU counterparts, suggesting that the combination of CNNs with LSTMs may mitigate the computational complexity typically associated with LSTM models.

3) Unidirectional vs. Bidirectional Models: Bidirectional models consistently outperformed their unidirectional counterparts, indicating that the ability to process sequences in both forward and backward directions is beneficial for capturing the full temporal context of lung sounds.

4) Model Complexity vs. Performance: While models with CNN layers generally performed better, the increased number of trainable parameters did not always guarantee superior performance. This is evident in the few instances where simpler models without CNNs outperformed their CNN-enhanced counterparts.

In conclusion, the integration of CNNs with RNNs, particularly in bidirectional configurations, provides a robust approach to lung sound classification. These models offer significant improvements in accuracy and reliability, making them valuable tools for respiratory disease diagnosis. Future work could explore optimizing the balance between model complexity and performance, particularly in tasks where simpler models sometimes outperform more complex ones.

V. DISCUSSION

This study aimed to develop an advanced hybrid deep learning model to enhance the detection and classification of pulmonary diseases by analyzing lung sounds. The integration of various neural network architectures, including LSTM, GRU, BiLSTM, BiGRU, and their CNN-enhanced counterparts, provided a comprehensive evaluation of model performance across different tasks. The results obtained from these models offer several insights into the effectiveness of these approaches and highlight key considerations for future research.

A. Impact of CNN Integration

The integration of Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) significantly improved model performance in most tasks. Specifically, models with CNN layers outperformed their non-CNN counterparts in 26 out of 32 comparisons, indicating the effectiveness of CNNs in capturing spatial features from spectrograms. The CNN layers effectively enhanced the models' ability to identify subtle patterns in the frequency domain of lung sounds, leading to higher F1 scores across various detection tasks.

However, the study also revealed that the addition of CNNs does not always guarantee superior performance. For instance, in the case of inhalation detection, the BiGRU model without

CNN achieved a slightly higher AUC (0.963) compared to its CNN-enhanced version (0.961). Similarly, GRU outperformed CNN-GRU in exhalation detection (AUC of 0.886 vs. 0.883), and BiGRU outperformed CNN-BiGRU for exhalation detection (AUC of 0.911 vs. 0.899). These findings suggest that while CNNs generally improve model performance, the added complexity may sometimes lead to diminishing returns, particularly in specific tasks where simpler models suffice.

B. Comparison Between LSTM and GRU Architectures

The comparison between LSTM and GRU models revealed that GRU architectures consistently outperformed LSTM models when CNNs were not included. This can be attributed to the GRU's simpler architecture, which makes it more efficient in learning temporal dependencies without overfitting. However, the performance gap between LSTM and GRU models narrowed when CNNs were added, as the CNN-LSTM models performed comparably to the CNN-GRU models. This indicates that the combination of CNNs with LSTMs helps mitigate some of the computational complexities typically associated with LSTM models, making them more competitive in terms of accuracy and efficiency.

C. Unidirectional vs. Bidirectional Models

The results demonstrated the superiority of bidirectional models (BiLSTM and BiGRU) over their unidirectional counterparts. The bidirectional models consistently achieved higher F1 scores, with improvements ranging from 0.4% to 9.8%. This performance boost is likely due to the bidirectional models' ability to capture information from both past and future contexts within the sequence, leading to a more comprehensive understanding of the temporal dependencies in lung sound data. The success of bidirectional architectures underscores the importance of capturing the full temporal context in time-series analysis, particularly in medical applications where precision is critical.

D. Model Complexity vs. Performance Trade-offs

While the integration of CNNs generally led to better performance, the increased number of trainable parameters did not always correlate with improved outcomes. The instances where simpler models without CNNs outperformed their CNN-enhanced counterparts highlight the importance of balancing model complexity with performance. This trade-off is particularly relevant in real-world applications, where computational resources and processing time are often limited. Therefore, future research should explore strategies to optimize model complexity without compromising accuracy, such as pruning techniques or the use of lightweight architectures.

E. Robustness and Generalizability of Models

The models incorporating CNN layers demonstrated flatter and lower MAPE curves across a wide range of threshold values, indicating greater robustness and consistency in performance. This robustness is crucial for the deployment of these models in clinical settings, where variability in data quality and recording conditions can significantly impact model performance. The ability of CNN-enhanced models to maintain stable performance across different thresholds suggests that they are better suited for practical applications, where varying conditions and noise levels are common.

F. Implications for Clinical Practice

The findings from this study have important implications for the use of deep learning models in the diagnosis of pulmonary diseases. The superior performance of models integrating CNNs and bidirectional RNNs indicates that these architectures are well-suited for the complex task of lung sound classification. The ability to accurately detect and classify adventitious lung sounds could lead to earlier and more accurate diagnoses of respiratory conditions, ultimately improving patient outcomes. Moreover, the use of automated, objective tools for lung sound analysis can reduce the variability and subjectivity associated with traditional auscultation, leading to more consistent and reliable diagnoses across different clinical settings.

VI. Future Research Directions

This study opens several promising avenues for future research. Firstly, integrating Temporal Convolution Networks (TCNs) with existing CNN-RNN architectures could enhance temporal modeling in lung sound analysis. Additionally, improving model interpretability, especially for complex CNN-LSTM and CNN-GRU hybrids, will be crucial for clinical adoption. Further exploration into multi-branch architectures and non-local operations may provide more robust and accurate models. Real-time implementation and deployment on resource-constrained devices, along with advanced data augmentation techniques, will help address practical challenges. Finally, cross-dataset validation, collaborative learning, and federated learning approaches should be prioritized to ensure models are generalizable, secure, and effective across diverse clinical settings.

References

Aykanat, M., Kılıç, Ö., Kurt, B. and Saryal, S., 2017. Classification of lung sounds using convolutional neural networks. *EURASIP Journal on Image and Video Processing*, 2017(1), pp.1-9. DOI: 10.1186/s13640-017-0213-2.

Demir, F., Sengur, A. and Bajaj, V., 2019. Convolutional neural networks based efficient approach for classification of lung diseases. *Health Information Science and Systems*, 8(1), pp.1-9. DOI: 10.1007/s13755-019-0091-3.

Hsu, F.-S., Huang, S.-R., Huang, C.-W., Huang, C.-J., Cheng, Y.-R., Chen, C.-C., Hsiao, J., Chen, C.-W., Chen, L.-C., Lai, Y.-C., Hsu, B.-F., Lin, N.-J., Tsai, W.-L., Wu, Y.-L., Tseng, T.-L., Tseng, C.-T., Chen, Y.-T. and Lai, F., 2020. Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database—HFLungV1. *PLOS ONE*, 15(12), pp.1-24. DOI: 10.1371/journal.pone.0243029.

Wang, X., Girshick, R., Gupta, A. and He, K., 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794-7803.

Ma, L.X., Zhang, H., Xu, Y. and Liu, B., 2020. LungRN+NL: An Improved Adventitious Lung Sound Classification Using Non-Local Block ResNet Neural Network with Mixup Data Augmentation. In *Proceedings of the Interspeech*, 2020. DOI: 10.21437/Interspeech.2020-2487.

Reyes, R.A.R., Hannuna, M.M., Markos, T., Nikolay, E.V. and Padilla, M.P., 2017. Automatic adventitious respiratory sound analysis: A systematic review. *PLOS ONE*, 12(5), pp.1-17. DOI: 10.1371/journal.pone.0177926.

Zhang, H., Cisse, M., Dauphin, Y.N. and Lopez-Paz, D., 2018. Mixup: Beyond empirical risk minimisation. In *International Conference on Learning Representations*, 2018.