Available online at www.dergipark.org.tr/en

INTERNATIONAL ADVANCED RESEARCHES and ENGINEERING JOURNAL

Journal homepage: www.dergipark.org.tr/en/pub/iarej

e-ISSN: 2618-575X

International Open Access

Volume 08
Issue 01

April, 2024

Research Article

Detection of wheeze sounds in respiratory disorders: a deep learning approach

Leen Hakki [a] ***and Gorkem Serbes*** [b,*]

[a]*Department of Biomedical Engineering, Yildiz Technical University, Istanbul, Turkey*
[b]*Department of Biomedical Engineering, Yildiz Technical University, Istanbul, Turkey*

ARTICLE INFO

Article history:
Received 10 December 2023
Accepted 14 April 2024
Published 20 April 2024

Keywords:
Convolutional Neural Networks
Deep Learning
Gated Recurrent Units
Long Short-Term Memory
Recurrent Neural Networks
Respiratory Sounds
Short Time Fourier Transform
Wheezes

ABSTRACT

Respiratory disorders, including chronic obstructive pulmonary disease (COPD) and asthma, are major causes of death globally. Early diagnosis of these conditions is essential for effective treatment. Auscultation of the lungs is the traditional diagnostic method, which has drawbacks such as subjectivity and susceptibility to environmental interference. To overcome these limitations, this study presents a novel approach for wheeze detection using deep learning methods. This approach includes the usage of artificial data created by employing the open ICBHI dataset with the aim of improving in generalization of learning models. Spectrograms that were obtained as the output of the Short-Time Fourier Transform analysis were employed in feature extraction. Two labeling approaches were used for model comparison. The first approach involved labeling after wheezing occurred, and the second approach assigned labels directly to the time steps where wheezing patterns were seen. Wheeze event detection was performed by constructing four RNN-based models (CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU). It was observed that labeling wheeze events directly resulted in more precise detection, with exceptional performance exhibited by the CNN-BiLSTM model. This approach demonstrates the potential for improving respiratory disorders diagnosis and hence leading to improved patient care.

1. Introduction

Among the leading causes of death worldwide are respiratory diseases, including chronic obstructive pulmonary disease (COPD), and asthma. COPD, the third leading cause of death globally, was responsible for 3.23 million deaths in 2019. In the same year, approximately 262 million people were affected by asthma, resulting in the deaths of 455,000 individuals [1-3]. The given statistics highlight the importance of early detection of respiratory diseases. The most frequently used method for respiratory disorders diagnosis is lung auscultation, which is a non-invasive and cost-effective method of diagnosis used to assess the condition of the lungs' health [4]. The auscultation method is used to listen to the sounds produced by the lungs. This method is essential for evaluating patients' respiratory symptoms (i.e. coughing, wheezing, or crackling). Although medical technologies for pulmonary diagnosis (i.e. spirometry, and chest X-ray) have advanced significantly, auscultation remains one of the commonly employed approaches for diagnosing respiratory sounds employing the traditional analog stethoscope [5]. Although auscultation with a stethoscope is valuable, it has some limitations, such as its subjectivity, as it relies on the expertise of the physician. Additionally, it provides inaccurate information when used in noisy environments. Besides, there is a risk of infection if it comes into direct contact with the patient [4]. Due to these limitations, researchers sought to improve the efficiency of auscultation by parameterizing lung sounds through computerized lung sound analysis or digital stethoscopes, which include sampling, filtering, feature identification, and lung sound classification [6].

Respiratory sounds are produced by the air moving through the lungs and airways during breathing [7]. The respiratory sounds can be categorized into two classes; normal and abnormal (adventitious) sounds. Normal respiratory sounds are characterized by a low noise in the inspiration phase and are difficult to hear during the exhalation phase. Their highest frequency range is under 100 Hz [7]. On the other hand, extra respiratory sounds that are not normally heard during breathing are called adventitious sounds. Those abnormal (adventitious sounds) can indicate the presence of pulmonary ailment [8]. The Internal Lung Sound Association has divided

* Corresponding author. Tel.: +90 533 965 81 89.
E-mail addresses: hakkileen@gmail.com *(L. Hakki)*, gserbes@yildiz.edu.tr *(G. Serbes)*
ORCID: 0009-0003-8155-8603 *(L. Hakki)*, 0000-0003-4591-7368 *(G. Serbes)*
DOI: 10.35860/iarej.1402462

adventitious sounds into continuous and discontinuous sounds. The adventitious sounds are further classified into wheeze and rhonchi, which are continuous sounds, and fine and coarse crackles, which are discontinuous sounds [9]. Crackles are discontinuous and explosive clicking or crackling sounds caused by the opening of small airways. The duration of crackles is short and usually less than 100 ms [9]. Crackles can indicate various health conditions such as pneumonia, chronic bronchitis, bronchiectasis, congestive heart failure, and obstructive pulmonary disease [10]. Wheezes are continuous sounds that occur as a result of air passing through narrow passageways due to blockage in the airways [9]. The duration of wheezes is considerably longer than the duration of crackles. Their duration lasts more than 100 ms, with an average of 250 ms, and they have a dominant frequency of 100 Hz or greater [9]. The sound of wheezes can differ among individuals and is influenced by factors such as the extent of the condition and the location where the stethoscope was positioned during auscultation [11]. Wheezes can indicate various health conditions such as asthma and bronchial stenosis [7]. A wheezing sound can be classified as monophonic or polyphonic. Monophonic wheeze refers to the wheezing sound heard consistently throughout the respiratory cycle with a uniform pitch. It is generally caused by airway narrowing because of a foreign body or tumor. Polyphonic wheezing involves multiple wheezes of different pitches occurring simultaneously, which is generally caused by asthma or COPD [7]. The presence of such different wheeze types makes wheeze detection more complex.

The human's ear ability to indicate adventitious respiratory sounds is limited. The intensity and type of abnormal sounds, as well as their amplitude, are significant factors contributing to detection errors in respiratory signal analysis. In addition, other environmental factors such as artifacts coming from patient movements, coughing, or speech are also limiting factors for lung auscultation. Hence, the validation of automatic adventitious sound detection algorithms should not solely rely on auscultation as the standard diagnosis method. Based on the reasons discussed, there is a need for the development of methods and algorithms that can accurately detect the adventitious sounds and implement them into smart stethoscopes for more effective diagnosis of respiratory illnesses. Therefore, in the last years, researchers put their efforts into developing classification models using machine learning-based algorithms. Before 2017, the available data on respiratory sounds were limited and insufficient [12]. Later, a challenge for lung sound classification was published at the International Conference on Biomedical and Health Informatics (ICBHI) 2017, containing 920 respiratory records collected from 126 subjects [13].

In this study, artificial data derived from the ICBHI dataset was generated to address issues like the scarcity of wheeze events and excessive noise in certain respiratory recordings. The spectrograms obtained from the short-time Fourier Transform (STFT) analysis of the recordings were utilized for feature extraction. Prior to inputting these spectrograms into the constructed model, two labeling approaches were employed. While prior studies mostly used convolutional neural networks (CNNs) for lung sound classification, this study employs a convolutional recurrent neural network (CRNN) architecture to better capture time-dependent patterns in the data. Detecting wheeze sounds not only indicates their presence but also their timing during breathing cycles. This research seeks to develop an efficient algorithm using digital auscultation recordings for improved respiratory disease detection with smart stethoscopes. While previous studies focused on the classification of respiratory sounds as will be discussed in the next section, this study aims to go beyond the classification by incorporating novel detection methods that provide detailed information on the occurrence, duration, and frequency of wheeze events accruing in a respiratory record. This detection method can allow for personalized treatment for individuals with respiratory disorders. Furthermore, integrating a detection algorithm into an electronic stethoscope provides real-time alerts of wheeze events present during medical examinations, resulting in more accurate diagnosis and improved care for patients with respiratory disorders.

The paper structure is as follows: Section 2 provides a comprehensive literature review of studies conducted using the ICBHI 2017 open-source data for adventitious sound classification. Section 3 details the methodology followed for building a wheeze detection model. Section 4 presents the results obtained, including the performance matrices and visual representations of them, in addition to a discussion of those results and a comparison between different models and labeling techniques. Finally, Section 5 encapsulates the conclusion of the study.

2. Literature Survey

Previous studies have employed different methods that have been applied for respiratory sound classification using the ICBHI 2017 dataset. In several studies, the ICBHI open dataset from 2017 has been a valuable resource [13]. Jakovljevic et al. applied a hidden Markov model to analyze Mel-frequency cepstral coefficients (MFCC) in respiratory sound data [14]. Chambres et al. took a different approach, combining low-level features and MFCC while using a decision tree to identify adventitious lung sounds, achieving an accuracy of 49.63% [15]. Kochetov et al. introduced a classification system for lung sounds, employing a noise-masking recurrent neural network (NMRNN) and utilizing STFT for feature extraction [16]. Ma et al. brought in spectral analysis techniques such as STFT and Wavelet Transform

(WT), implementing a bilinear ResNet (bi-ResNet) neural network to classify respiratory sounds with an accuracy of 52.79% [17]. Ngo et al. ventured into the use of Gamatongue spectrograms, employing an ensemble of Clustering deep neural networks (C-DNN) and Autoencoder networks to classify respiratory cycle anomalies [18]. Acharya et al. proposed a hybrid CNN-RNN model for classifying features derived from Mel spectrograms [19]. Serbes et al. used STFT and WT for feature extraction, employing a support vector machine (SVM) as their classifier, resulting in an accuracy of 49.86% [20]. Demir et al. used the STFT for feature extraction and presented two different deep-learning approaches. The first method included a 16-layer deep convolutional neural network (VGG-16 CNN) with an SVM classifier, achieving an accuracy of 65.5%. In the second approach, transfer learning was employed with the CNN model and a softmax function, yielding an accuracy of 63.09% [21]. In 2020, Demir et al. introduced a CNN model trained using spectrogram images paired with a Linear Discriminant Analysis (LDA) classifier and the Random Subspace Ensembles (RSE) method, resulting in an accuracy of 71.15% [22]. Bilal M. devised a system focused on extracting spectrograms from lung sound signals and feeding them into a custom 12-layer CNN, achieving an accuracy of 64.5% [23].

Asatani et al. employed STFT for extracting the features from the data and integrated Convolutional Recurrent Neural Networks (CRNNs) with bidirectional Long Short-Term Memory (bi-LSTM) blocks to enhance classification accuracy [24]. Similarly, Yang et al. employed STFT for feature extraction and developed a framework using ResNet-18, incorporating Squeeze-and-Excitation (SE) and Spatial Attention Blocks (SA) [25]. In a related vein, Liu et al. classified adventitious respiratory sounds, using Log Mel-filterbank (LMFB) for feature extraction and employing CNN for classification, achieving an accuracy of 81.62% [26]. Likewise, Perna and Tagarelli used MFCCs for feature extraction and recurrent neural networks (RNN) for adventitious respiratory sound classification [27]. Zulfiqar et al. used spectrograms with Artificial Noise Addition (ANA) for feature extraction, employing various CNN architectures (AlexNet, ResNet50, VGG16, and Baseline) for the classification of diverse adventitious respiratory sounds [28]. Similarly, Nguyen and Pernkopf incorporated STFT and Log-Mel for feature extraction, using a pre-trained ResNet model to classify adventitious lung sounds [29]. Ntalampiras and Potamitis introduced a unique feature set based on wavelet analysis, implementing a directed acyclic graph (DAG) network architecture comprising hidden Markov models (HMM) to model their distribution [31]. So far, most research concerning ICBHI data has focused on classifying respiratory sounds. This involves training a machine learning model to identify the presence or absence of certain classes but does not provide details about where these classes are located. Detection combines classification and localization, offering information about the type of object present and its specific location. In our previous study [34], we developed a method for wheeze detection using STFT for feature extraction, and CNN-GRU deep learning model. The built wheeze detection model achieved an F1 score of 0.73. This study is built upon the previous work, however, in this study, an alternative labeling method and other RNN architectures were employed to improve wheeze detection.

Wheeze detection offers unique advantages compared to sound classification alone. In addition to identifying the wheezing event occurrence, this method predicts the duration and frequency of wheezing events throughout the diagnostic process. This leads to tailored treatments and individualized care, enabling healthcare providers to customize interventions and medications for individuals.

3. Materials and methods

In this section, the methods employed for constructing a detection model are discussed. Figure 1. provides a visual representation of the methodology workflow.

3.1 Data Description

For this study, the public dataset of the 2017 ICBHI competition was used, containing a total of 920 lung sounds recorded by two research teams in Portugal and Greece. The recordings ranged in length from 10 to 90 seconds with sampling frequencies of 4 kHz, 10 kHz, and 44.1 kHz. The cycles can be classified as crackle, wheeze, both crackle, and wheeze, or no ambient noise [13]. Wheezes and crackles are annotated in the dataset with detailed start and end times. The first column in the detailed annotation indicates the start time, the second indicates the end time, and the third indicates the name. Text files with detailed events were used for the study.

3.2 Signal Processing

The ICBHI database contains recordings with sampling frequencies of 4 kHz, 10 kHz, and 44.1 kHz. To maintain consistency and compatibility across audio files, they were resampled to a 4 kHz sampling rate. Afterward, to reduce unwanted noise arising from different sources like coughing, intestinal/cardiac sounds, and stethoscope movement, a 12th-order Butterworth band-pass filter with cutoff frequencies set at 120 and 1800 Hz was applied to the recordings.

3.3 Preparing the Data for Training

According to the annotations, among a total of 920 audio files, only 341 of them contain one or more instances of wheezing events, totaling 1879 wheeze events in those files. The duration of wheeze segments varies between 0.03 seconds and 5.80 seconds.

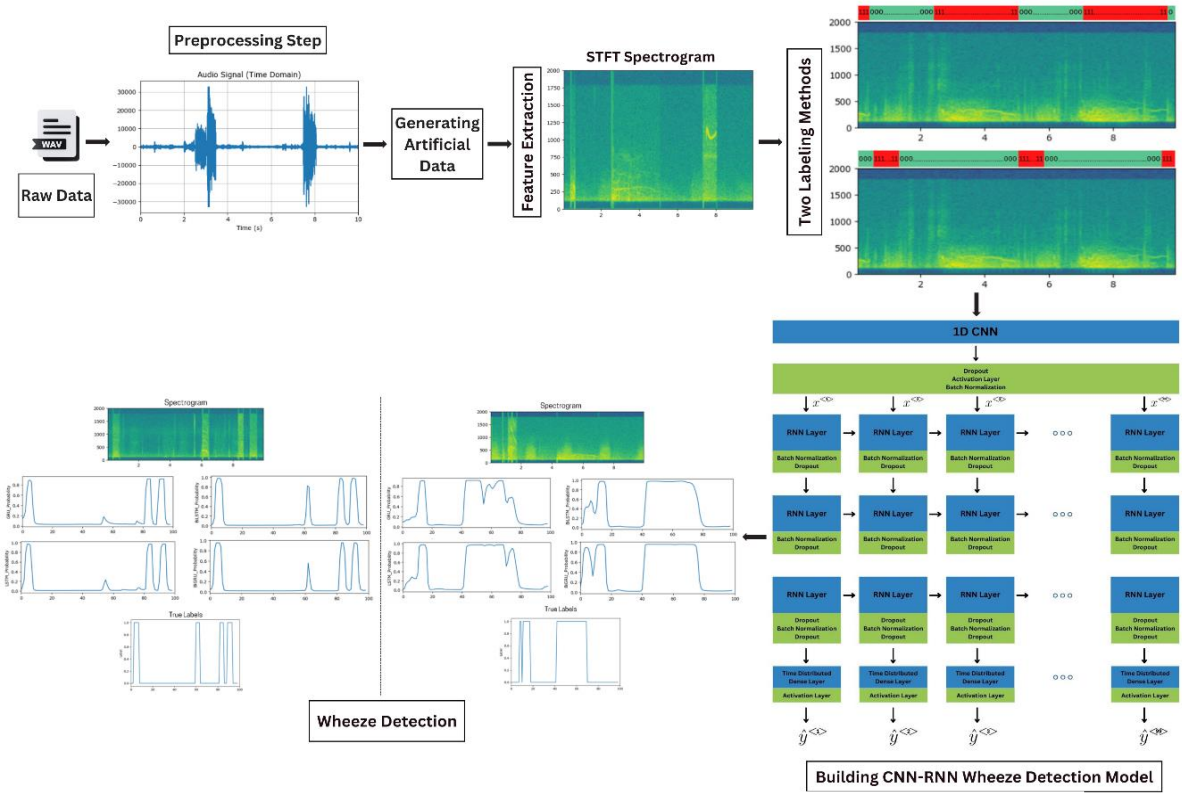

Figure 1. The flowchart of detection framework. The figure contains a visualization of the CRNN model used in the detection process

Figure 2. shows the histogram of the durations, where it can be seen that the majority of wheezes are between the interval of 0.1 seconds and 0.4 seconds. The range of the highest bar, which represents the most frequent durations, in the histogram is between 0.17 and 0.31 seconds. Those numbers match with the literature, which states that wheeze sounds last more than 100 ms on average.

The used dataset encompasses approximately 5.5 hours of recordings as mentioned before. However, when considering the wheezes specifically, the total duration amounts to only approximately 18 minutes of the data. Therefore, the duration of wheezing is only 5.61% of the overall duration of the provided data. This scarcity of wheezing occurrences in the dataset presents a considerable obstacle to developing a reliable wheeze detection model. In addition, during data review, inaccurate event labels in the annotation file were noticed, also, some audio files remained heavily distorted by excessive noise, even after applying the bandpass filter, leading to the emergence of oscillatory patterns that exhibit characteristics similar to wheezing, i.e., exhibiting characteristics like speech or motion artifacts. Consequently, the model may recognize them as instances of wheezing, or it can categorize wheezing as a normal sound. To overcome these limitations, an artificial data generation technique was employed. By creating synthetic data using the annotation files, it is possible to expand the existing dataset and obtain a larger and more diverse set of wheeze events with reduced noise. This strategy aims to enhance the precision of the models by furnishing a more comprehensive training dataset.

3.4 Generating Artificial Data

The data was initially split into training and testing sets with a 70% to 30% ratio. Wheezing-containing audio files were identified, and wheezing segments were extracted from them for both training and testing datasets. Figure 3. illustrates the waveform of one of the wheeze segments, where the oscillatory behavior of the wheeze can be clearly seen.

Audio files without wheeze events were also identified, and their first 10 seconds were downloaded. Before using them as background for synthetic data, each file underwent a review, with those containing talking or excessive ambient sounds excluded.

Using wheeze segments and background audio, synthetic data was created. The code randomly selected 10-second backgrounds and added 2 to 4 wheezes to each. 130 different

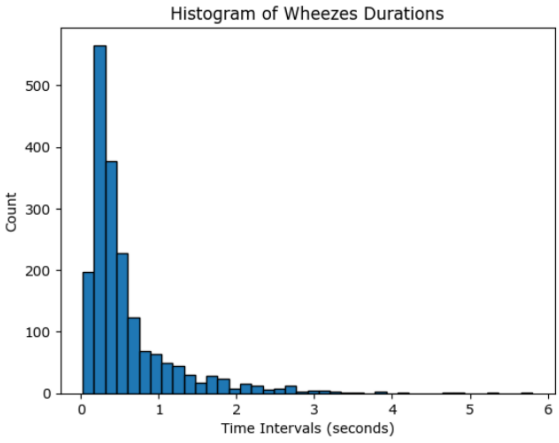

Figure 2. The Histogram of Wheeze Durations

backgrounds with lower noise levels and minimal talking sounds were selected. Among these, 100 background signals were employed for training, while 30 background signals were used for testing.

The code ensures no overlapping segments in selected audio clips to avoid simultaneous wheeze events. It creates 2, 3, or 4 wheeze segments in 10-second sound files. Having at least two wheeze events per audio is crucial to balance the skewed labels (mostly 0s), reducing the risk of overfitting. This approach generates 1500 training and 200 testing audio files.

3.5 Short-Time Fourier Transform Spectrogram

The Short-Time Fourier Transform (STFT) is a commonly used method for the time-frequency analysis attributes of a signal. It involves dividing the input signal into overlapping windows and subsequently applying the Fast Fourier Transform (FFT) to each of these windows. The STFT shows how the frequency content of the signal changes over time, making it a powerful tool for analyzing time-varying spectral features. The mathematical representation of the STFT, X[*m*,*w*], is as follows [32]:

$$X[m,w] = \sum_{n=0}^{N-1} x[n]w[n-m]e^{-iwn} \quad (1)$$

Where in Equation (1), $x[n]$ represents the input signal, $w[n]$ refers to the window function and N is the length of the FFT. Squaring the magnitude of the STFT, $|X[m,w]|^2$, gives the spectrogram which is the visual representation of the signal in the time-frequency domain.

When visualizing the audio signal's spectrogram, it becomes evident that specific patterns associated with the characteristic frequency range become clearly visible. For the best visualization of the patterns, and for obtaining accurate features to be fed to the model, the parameters of the STFT should be selected carefully by giving attention to both the time resolution and the frequency resolution of the spectrogram. A narrow window provides good time resolution but sacrifices frequency resolution, while a wide window offers poor time resolution but good frequency resolution. Since wheezes last a long time and have a narrow frequency pattern, there is a need for better frequency resolution to visualize them in the time-frequency domain.

Figure 4 displays a plot of an audio signal alongside its corresponding spectrogram. In this study, a Blackman-Harris (BH) window was used for wheeze visualization with a length of 512 and a 192 overlap (75% of the window size). This allows us to clearly see wheezes as narrowband spectral patterns, often referred to as "snakes" in the spectrogram.

3.6 Labeling the Data

The labels were produced using two distinct approaches. In the first approach, the labels for the audio signals were synchronized with the time frames of the spectrograms, assigning one label for every three-time steps. The portions

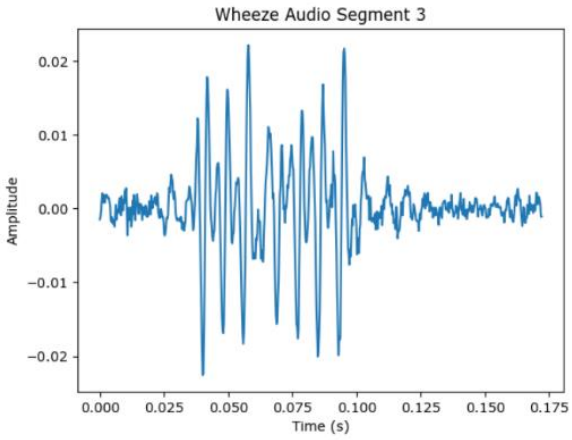

Figure 3. Wheeze segment plot in the time domain

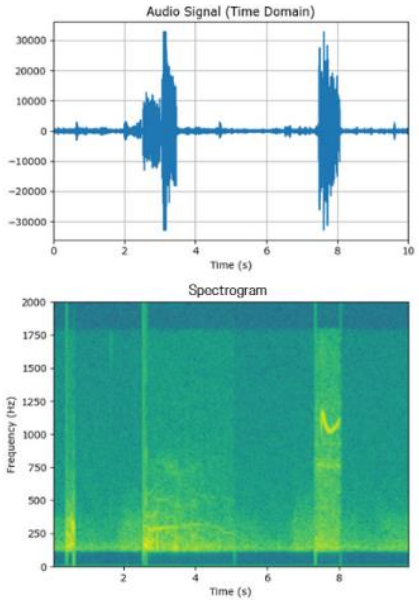

Figure 4. Audio signal plot in the time domain and its corresponding STFT spectrogram

of the audio that lack wheezes are assigned zero. While labeling the portions of wheezing, a specific approach was employed.

In this approach, a target label of 1 is assigned to the 10 consecutive time steps following the end of a wheeze-containing clip. The goal of labeling in this way is to train the model to recognize when wheezing sounds show up, with a focus on the time immediately after the clip ends. Each generated spectrogram has a label in the shape of (1, 99) which is the output of the model. The labeling of segments is visually represented in Figure 5A, where a binary classification is applied: "the time steps immediately after the wheezing event shows up" (labeled as 1) or " time steps without wheezes, or wheezing events" (labeled as 0).

In the second approach, the labels are assigned directly to the time steps where the wheeze occurs. In this way, the model is supposed to be able to predict the exact time of the

wheeze appearance. Similar to the first approach, the shape of the labels is (1, 99). The labeling of segments is visually represented in Figure 5B, where a binary classification is applied: "with wheeze" (labeled as 1) or "without wheeze" (labeled as 0).

3.6 RNN-Based Model Architecture

In this study, four RNN-based models, namely CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU, were investigated. The built model architecture includes a single convolutional layer, followed by three RNN layers, and finalized with a dense layer.

The model starts with a convolutional layer, taking a 309-time step input. The convolutional layer is critical to extract low-level features and decrease the output dimensionality. This layer has 196 units, uses a 15-sized kernel, and has a stride of 3. This adjustment aligns the model's output dimensionality with the label dimensions (99). The CNN layer accelerates learning by reducing input dimensionality. Consequently, batch normalization is applied to normalize the feature, and then the ReLU activation function is used to introduce non-linearity. A dropout layer is added to prevent overfitting.

Following the CNN layer, RNN layers are employed. The architecture of the four models is identical in terms of the number of layers, dropouts, batch normalizations, and other parameters. The only difference lies in the type of RNN blocks used. The first layer utilizes 512 units and return sequences, allowing the model to capture temporal relationships and patterns within the dataset. Dropout and batch normalization are applied to enhance the stability of the model. The next RNN layer also utilizes 512 units. This layer further captures complex temporal relationships in the data. Dropout and batch normalization are applied similarly to the previous layer. The last RNN layer employs 256 units and serves to prepare the data for the subsequent output layer. Dropout and batch normalization are applied again to ensure robustness and prevent overfitting. The final layer consists of a time-distributed dense layer with sigmoid activation. The model's architecture is presented in Table 1.

The model was trained for 30 epochs. For the evaluation of the models, several metrics were generated to assess their performance. These metrics include accuracy, precision, recall, and F1 score.

4. Results and Discussion

4.1 Models' Accuracies

Table 2 presents the evaluation metrics for the model, comparing the two different labeling techniques. The CRNN models with positive labels starting after the events achieved remarkably similar accuracies ranging between 0.82 and 0.84. Notably, the GRU-based model achieved the lowest accuracy, while the bidirectional models exhibited the highest accuracy among the four models. These accuracies indicate that the model can correctly identify wheeze patterns in 82-84% of cases. Similarly, the CRNN models with the positive labels aligned directly on the wheeze events demonstrated accuracies within the same range as the models using the alternative labeling technique. However, in this case, the LSTM-based model achieved the lowest accuracy of 0.82.

While accuracy is a crucial metric for classification tasks, in scenarios where the distribution of labels is heavily skewed towards the "0" class, accuracy may not provide the most informative assessment. In such cases, defining more meaningful metrics such as F1 score, Precision, and Recall, offers a more insightful evaluation of the model's performance. By considering those metrics, a better understanding of the model's ability to correctly predict positive class while minimizing false positives and false negatives can be achieved.

4.2 Models' Precision and Recall Values

Precision evaluates how effectively the model can correctly identify positive predictions. Equation (2) represents the mathematical calculation of precision. A higher precision score signifies the model's capacity to correctly identify wheezing patterns while reducing the misclassification of non-wheezing events as wheezing.

$$Precision = \frac{\sum TP}{\sum(TP + FP)} \tag{2}$$

Recall assesses the model's capacity to accurately identify all positive instances. Equation (3) represents the mathematical calculation of the recall. The recall score evaluates the model's ability to detect the real instances of wheezing in the dataset, ensuring that it doesn't overlook any existing wheezing events.

$$Recall = \frac{\sum TP}{\sum(TP + FN)} \tag{3}$$

Table 1. Model Architecture

Layer Type	Kernel Attribute	Activation
Conv1D	15 (196 Filters)	ReLu
BatchNormalization		-
Dropout	0.8	-
RNN	512 units	-
Dropout	0.8	-
BatchNormalization		-
RNN	512 units	-
Dropout	0.8	-
BatchNormalization		-
RNN	256 units	-
Dropout	0.8	-
BatchNormalization		-
Dropout	0.8	-
TimeDistributed	Dense 1	Sigmoid

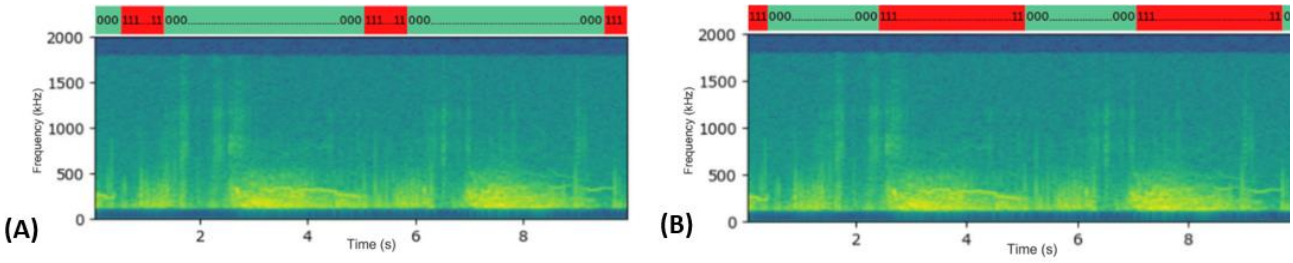

Figure 5. Visual Representation of the Labeling. (A) Positive Labeling After Wheeze Events and (B) Positive Labeling Directly on Wheeze Events

Table 2. Accuracy, Precision, Recall, and F1 Score Values

		Positive Labels After Wheeze Events				**Positive Labels on Wheeze Events**			
Metric		**CNN-LSTM**	**CNN-GRU**	**CNN-BLSTM**	**CNN-BGRU**	**CNN-LSTM**	**CNN-GRU**	**CNN-BLSTM**	**CNN-BGRU**
Accuracy		0.83	0.82	0.84	0.84	0.82	0.83	0.84	0.84
Precision	**0**	0.86	0.87	0.87	0.85	0.85	0.85	0.86	0.86
	1	0.75	0.60	0.67	0.78	0.73	0.75	0.75	0.79
Recall	**0**	0.95	0.90	0.92	0.96	0.92	0.93	0.92	0.94
	1	0.47	0.54	0.53	0.43	0.57	0.56	0.61	0.58
F1 score	**0**	0.89	0.88	0.90	0.91	0.88	0.89	0.89	0.90
	1	0.58	0.57	0.59	0.56	0.64	0.64	0.68	0.67
Macro average	**Precision**	0.80	0.74	0.77	0.82	0.79	0.80	0.81	0.82
	Recall	0.71	0.72	0.73	0.70	0.74	0.74	0.77	0.76
	F1 Score	0.74	0.73	0.74	0.73	0.76	0.76	0.78	0.78

Examining the precision values in Table 2, it becomes evident that almost all models supplied with positive labels after wheeze events achieve relatively high values for negative class (0) which refers to wheeze absence, ranging from 0.86 to 0.87. This indicates the ability of the models to correctly classify non-wheezing events. However, the precision for the positive classes (1) which refers to the wheeze events, varies between the models. The CNN-BiGRU model achieves the highest precision of 0.78, while the CNN-GRU model has the lowest precision of 0.60.

The macro average represents the average across both classes, treating them equally. The macro precision average ranges between 0.74 and 0.82, with the highest value for the CNN-BiGRU model. These results indicate the models can precisely classify approximately 74%-82% of the positive instances. Regarding recall of the same labeling technique, the models show varying performances. For the negative class (0), the recall values range between 0.90 and 0.96, suggesting the models' effectiveness in capturing true negative (TN) instances.

However, for the positive class (1), the recall values range from 0.43 to 0.54, indicating that the model's ability to capture positive wheeze events ranges between 43% and 54%. The macro average of the models ranges between 0.70 to 0.73, with the highest value for the CNN-BiLSTM model. This indicates that the models exhibit a balanced

performance in terms of correctly identifying wheezing events and non-wheezing events.

When examining the precision and recall values derived from the second labeling technique, which directly labels wheeze events as positive, positive differences in performance become apparent when compared to the first labeling technique within the same table. In terms of precision, all the models achieve relatively high values for both the negative class (0) and the positive class (1) when compared with the alternative labeling technique. The precision value values range between 0.85 and 0.86 for the negative class and from 0.73 to 0.79 for the positive class.

For recall, the models again exhibit varying performance with relatively high values for the negative class. However, the recall values for the positive class, in the second technique that directly labels wheeze events as positive, range from 0.56 to 0.61, suggesting that the models have varying degrees of success in identifying wheezing events. Overall, the recall values are relatively higher than the alternative model. The macro average of the models ranges between 0.74 to 0.77, with the highest value for the CNN-BiLSTM model.

4.3 Models' F1 Scores

The F1 score is calculated as the harmonic mean of precision and recall. It is a metric that combines precision and recalls into a single value providing a balanced measure of models' performance. Considering the unbalanced nature of the labels, where the positive class (wheeze events) is a minority class, it is crucial to focus on the F1 score of the positive class (1). Looking at Table 2, it can be observed that among the models, CNN-BiLSTM and CNN-BiGRU consistently demonstrate higher F1 scores for wheeze events across both labeling techniques. For positive labeling after wheeze events, the BiLSTM model achieves the highest F1 score of 0.59 for the positive labels, while for labeling directly on wheeze events both bidirectional models give the same highest F1 score of 0.68. Comparing the F1 scores for the positive class, the second labeling technique outperforms the models with positive labeling after the wheeze events. This suggests that labeling directly on wheeze events provides a more balanced correct identification of wheeze events (precision) and captures all actual wheeze events (recall).

In order to provide a more comprehensive overview of the model's performance, two bar charts were illustrated to visualize the accuracy and the macro averages of the precision, recall, and F1 score. Figure 6 shows the scores of the models with positive labels after wheeze events, while Figure 7 depicts the scores of the models with positive labels on wheeze events. These charts provide a concise and visual representation of the data presented in Table 2. Upon closer inspection of the figures, it becomes evident that the bidirectional models outperform other architectures in both labeling techniques.

4.4 Visualization of the Obtained Results

In this study, results obtained from the models are visualized to offer a visual assessment of the model's performance in detecting wheezing events. Both the spectrogram and the prediction values were used to assess the model's ability in this regard. In this section, two examples are presented. The first example is from the artificial test data, and the second is from one of the original audios of ICBHI data. Figure 8 illustrates the results of one created audio example with positive labeling after the wheeze event. Peaks or changes in the prediction values of the probability graphic show when the model predicts the occurrence of a wheeze event. It can be seen that all the models except the BiGRU model could predict all the wheezes in the audio. GRU model's prediction values were less than 0.8, whereas the LSTM-based models gave prediction values of more than 0.9. Figure 9 illustrates the results of the same example but with the positive labels aligned directly on the wheezes. Looking at the true labels it can be seen that the bidirectional models could identify all the present wheeze events in the audio.

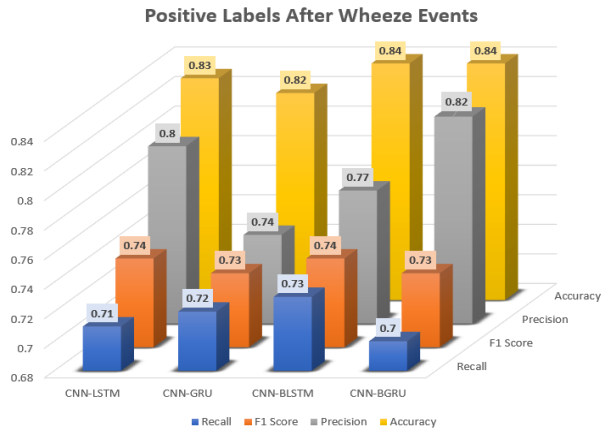

Figure 6. Accuracy, precision, F1 score, and recall scores for models with positive labels after wheeze events

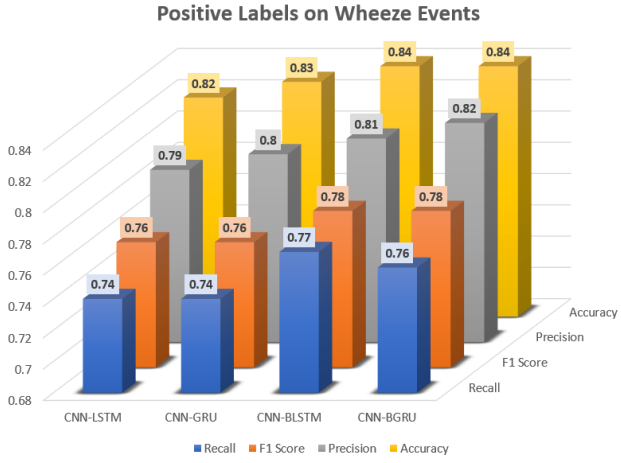

Figure 7. Accuracy, precision, F1 score, and recall scores for models with positive labels on wheeze events

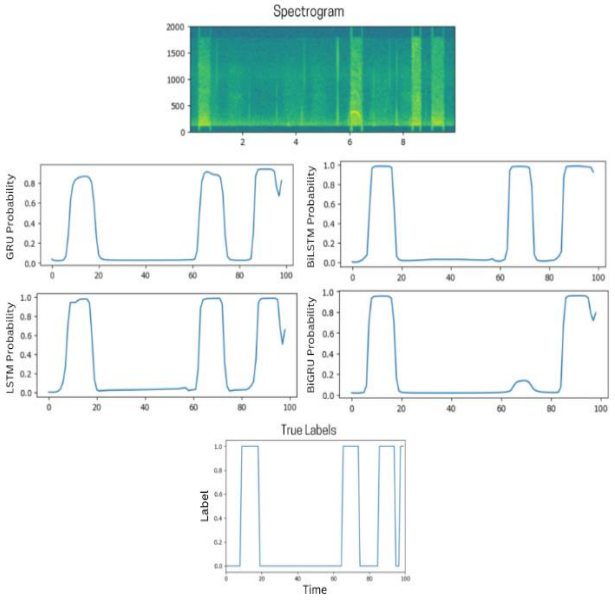

Figure 8. Visualization of the spectrogram, the predictions, and the true labels with positive labels after the wheeze events.

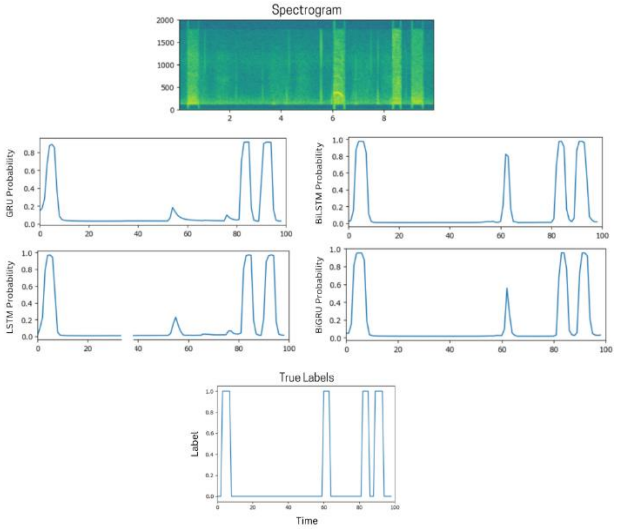

Figure 9. Visualization of the spectrogram, the predictions, and the true labels with positive labels aligned on the wheeze events

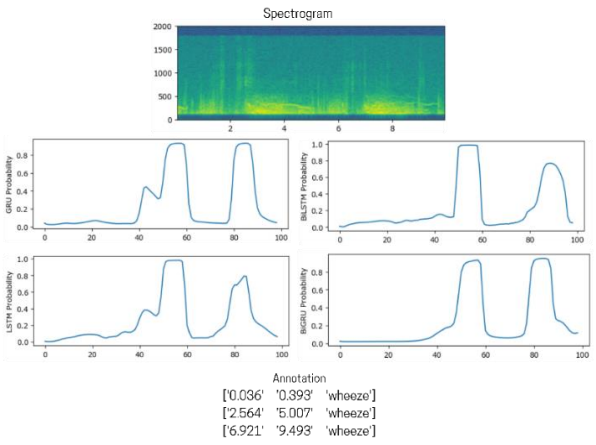

Figure 10. Visualization of the spectrogram, the predictions, and the true labels with positive labels after the wheeze events

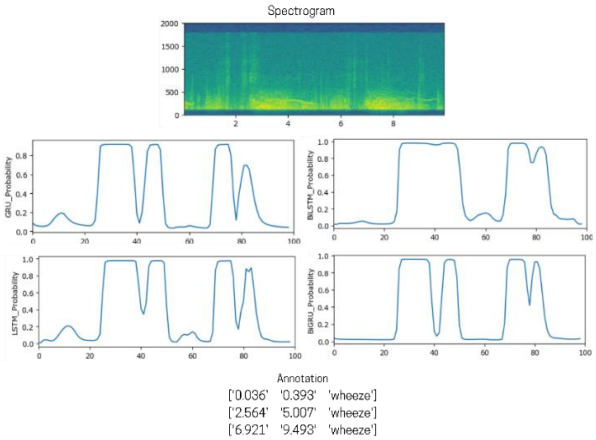

Figure 11. Visualization of the spectrogram, the predictions, and the true labels with positive labels aligned on the wheeze events.

The BiGRU model could detect the wheeze after the second 6 with a prediction value of 0.6, whereas the BiLSTM model was able to detect the same wheeze event with a prediction value of 0.8. On the other hand, the unidirectional models were not able to detect it. Figure 10 illustrates the results of one of the original audio examples with positive labeling after the wheeze events. From the annotation, it can be seen that all the models were not able to detect the short wheeze event at the duration between 0.036 and 0.393, whereas all of them were able to predict the other two events. Figure 11 illustrates the results of the same audio, but with labels directly assigned to the wheeze events. Looking at the annotations, it can be noticed that all models were not able to predict the same wheeze event at the start of the audio, however, for the other wheeze events the bidirectional models gave a more accurate prediction, with the best prediction given from the BiLSTM model. A set of other synthetic data and original data were also visualized, and as a visual result, it was noticed that the models with positive labels assigned directly to the wheeze events, especially the BiLSTM model, were able to detect the wheeze events more precisely.

4.5 Comparison Between Models

The conclusive comparison between LSTM-based and GRU-based models is difficult, however, in terms of computation times, the GRU-based models have been proven to be faster than the LSTM models [33]. As mentioned before, with the labels being unbalanced, and the positive class (wheeze events) being a minority, it is crucial to focus on the F1 score of the positive class (1). Therefore, focusing on the positive class of the F1 score it can be noticed that the LSTM-based models outperform other models in the first labeling technique. This difference in performance could be explained by the LSTM architecture, which includes a memory cell and gating mechanisms that allow it

to capture and remember long-term dependencies and retain information over a longer sequence. In the case of wheeze detection, where multiple time steps may be required to detect wheezing, these long-term dependencies are particularly useful. TN and TP values shown previously may be slightly lower in GRU models due to their relatively weaker ability to retain and utilize longer-term information.

On the other hand, in positive labeling directly on wheeze events, looking at the F1 scores of positive labels, it can be noticed that the bidirectional models tend to outperform the unidirectional models. Bidirectional models have the advantage of considering both past and future contexts during training. The reason for this can be that bidirectional models are capable of capturing bidirectional dependencies between wheezing events. Using bidirectional models, input sequences can be processed in both forward and backward directions, providing a more comprehensive input understanding.

Looking at the tables above, it is noticed that in both labeling techniques, the CNN-BiLSTM model achieves the highest F1 score for positive labels and the macro average. This can be attributed to the advantages offered by its architecture which combines the strength of both LSTM and bidirectional models. Therefore, the CNN-BiLSTM model with its combination of long-term dependencies capture and remember from LSTM and bidirectional context modeling, offers a favorable balance between capturing temporal patterns of wheeze events and incorporating bidirectional dependencies, making it well-suited for the wheeze detection task.

4.6 Comparison Between Labeling Techniques

In this study, two different labeling techniques were used. The results of each technique were obtained separately and presented in previous parts. The first labeling approach involves assigning positive labels following wheeze events. By applying this labeling method, the models consider the entire context of the wheeze events' pattern, allowing for a thorough analysis of those patterns. This labeling approach is generally applied in speech recognition tasks, where the labeling is performed after the speech utterance. However, in contrast to speech recognition or trigger word detection scenarios, the duration of spoken words is typically constant, whereas the lengths of wheeze events in the dataset vary significantly. This variability in wheeze duration presents a challenge in labeling wheezes after the event, as a fixed number of positive labels may not adequately capture the varying lengths of wheezes present in the data set. This challenge suggests relatively lower F1 scores for positive labels compared with the alternative method. Implementing this labeling technique along with using the CNN-LSTM model for wheeze detection offers the advantage of real-time alarm generation. By implementing the algorithm into a stethoscope, it may be possible to receive an instant alert when a wheeze occurs. This instantaneous feedback can be valuable in medical settings to tackle the problems mentioned in the introduction section of this study. However, using bidirectional models with this labeling method requires waiting for the recording of respiratory sounds to finish, which limits the real-time capabilities.

Positive labeling directly aligned with the wheeze event offers several advantages. Firstly, more precise identification of wheeze occurrence, allowing for accurate duration estimation, can be achieved. Furthermore, direct labeling in the event can capture fine-grained details of the wheezing pattern, leading to a higher F1 score for positive labels. It enables the model to learn specific wheeze features, resulting in enhanced differentiation between wheeze and non-wheeze segments. To implement the CNN-LSTM system with such a labeling approach to an electronic stethoscope for wheeze detection to provide an instant alarm, it typically needs to wait for the wheeze pattern to reach a recognizable state. While the system may not be able to provide an instant alarm at the exact onset of the wheeze, once a sufficient portion of the wheeze event is detected and recognized, the system can trigger an alarm to indicate the presence of a wheeze.

In summary, even though the first labeling technique allows comprehensive analysis of the wheeze pattern, the variable lengths of wheeze events can be challenging and result in lower F1 scores for positive labels. On the other hand, labeling directly aligned with the event provides precise identification and duration estimation of wheeze occurrence, yielding higher F1 scores.

4.7 Comparison with Other Studies

Table 3 shows some of the most recent and relevant studies in the literature for the ICBHI 2017 database. The table contains the studies that utilize STFT or comparable time-frequency representations, such as Mel spectrograms or MFCCs, and employ deep learning-based methods for classification. In the table, the column "parameters" denotes the window length used for time-frequency representation. Furthermore, columns "Sen", "Spe", "Sco", and "Acc" refer to the ICBHI 2017 benchmark sensitivity, specificity, scoring result, and classification accuracy, respectively. Although the studies listed in Table 3 employed methods similar to those utilized in this study, it's crucial to note a distinction. These studies focused on classifying abnormal respiratory sounds, and training models to identify the presence or absence of adventitious lung sounds in the record. However, they lacked the capacity to provide localization information about these classes. As mentioned previously, the majority of works on the ICBHI 2017 database have centered around classifying respiratory sounds without offering details about where these sounds occur. In contrast, our study stands out by aiming to develop a model for wheeze detection, which integrates both the classification and the localization of the wheeze sounds.

Table 3. Studies in Literature for the ICBHI 2017 Dataset. Column "parameters" denotes the window length for employed time-frequency representation.

Reference	Time-frequency representation		Learning model	Results (%)			
	Type	parameters		Sen	Spe	Sco	Acc
Kochetov et al. [16]	STFT	500 ms	RNN	58.4	73	65.7	-
Acharya et al. [19]	Mel spectrograms	60 ms	hybrid CNN-RNN	-	58.01	-	-
Liu et al. [26]	LMFB	-	CNN	-	-	-	81.62
Asatani et al. [24]	STFT	40 ms	CRNN, bi-LSTM	63	83	73	-
Perna &Tagarelli [27]	MFCCs	250 ms	RNN	64	84	74	-
Saraiva et al. [30]	MFCCs	-	CNN	-	-	-	74.3

5. Conclusions

Chronic respiratory disorders have a significant impact on lung function and overall health, with conditions such as chronic obstructive pulmonary disease and asthma contributing to a high number of deaths globally. Early diagnosis and regular monitoring of respiratory illnesses are crucial for effective management. While traditional auscultation using a stethoscope has been a common diagnostic tool, it has limitations in terms of subjectivity and accuracy. Computer-based respiratory sound analysis has emerged as a promising approach to improve the objectivity and efficiency of diagnosis.

The objective of this study is to develop a computer-based system capable of detecting wheeze sounds, to address the limitations of the traditional auscultation devices mentioned before. By implementing computer-based respiratory sound analysis using artificial data derived from ICBHI open data, conventional recurrent (CRNN) models were trained to detect wheeze in the created respiratory recordings.

In the implementation of the wheeze detection model, one of the major challenges faced was the limited open data of respiratory sound recordings. The available open data that was used in this project is relatively small, comprising only 920 audio files, of which only 341 files contain one or more wheeze events. In this study, to tackle this limitation, artificial data was generated from the ICBHI open data. However, having a larger dataset with a higher number of patients, a higher number of wheeze events, and reduced noise would significantly enhance the performance of the wheeze detection models, leading to more efficient and accurate wheeze detection capabilities.

Before training the data, two different methods of labeling were applied to the created recordings, each with its advantages and considerations. Those techniques are; positive labeling after wheeze events, and positive labeling aligned with the wheeze events. The second labeling method showed more precise identification, duration estimation, and enhanced differentiation between wheeze and non-wheeze segments.

The comparison between different models and labeling techniques showed that the LSTM-based models demonstrated better performance in terms of F1 score for positive labels after the wheeze events. On the other hand, the bidirectional models showed superior performance and F1 scores for positive labeling directly aligned with wheeze events. The CNN-BiLSTM model emerged as the most effective model, leveraging the strengths of both LSTM and bidirectional architecture. It combines the ability to capture long-term dependencies and incorporate bidirectional dependencies making it suitable for wheeze detection problems.

In future studies, it is recommended to address the limited dataset issue by acquiring a larger dataset with more patients, wheezing events, and reduced noise. Moreover, using K-fold cross-validation can lead to a more robust analysis, enhancing the reliability of wheeze detection models. This would further improve the efficiency and accuracy of computer-based respiratory sound analysis in order to foster their potential for early diagnostics and monitoring of respiratory diseases.

Declaration

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The author(s) also declared that this article is original, was prepared in accordance with international publication and research ethics, and ethical committee permission or any special permission is not required.

Author Contributions

L. Hakki and G. Serbes developed the methodology, performed the analysis, and wrote the manuscript together. G. Serbes proofread the manuscript.

Nomenclature

ICBHI : International Conference on Biomedical and Health Informatics
CRDs : Chronic Respiratory Disorders
COPD : Chronic Obstructive Pulmonary Disease
CNN : Convolutional Neural Network
RNN : Recurrent Neural Network
NMRNN: Noise Masking Recurrent Neural Network
ANA : Artificial Noise Addition
HMM : Hidden Markov Model
SVM : Support Vector Machine
DAG : Directed Acyclic Graph
LDA : Linear Discriminant Analysis
RSA : Random Subspace Ensembles
SE : Squeeze-and-Excitation
SA : Spatial Attention Block
STFT : Short-Time Fourier Transform
WT : Wavelet Transform
LMFB : Log Mel-Filterbank
MFCC : Mel-Frequency Cepstrum Coefficient
GRU : Gated Recurrent Unit
LSTM : Long Short-Term Memory
Bi-LSTM: Bidirectional Long Short-Term Memory
Bi-GRU : Bidirectional Gated Recurrent Unit
TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative
Hz : Hertz
ms : milliseconds

References

1. Cukic, V., Lovre, V., Dragisic, D., & Ustamujic, A. *Asthma and chronic obstructive pulmonary disease (COPD) – differences and similarities.* Materia Socio-Medica, 2012. **24**(2): p. 100.
2. World Health Organization. (n.d.). Chronic obstructive pulmonary disease (COPD). World Health Organization. Retrieved [cited October 25, 2022]; Available from: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).
3. Liang, R., Feng, X., Shi, D., Yang, M., Yu, L., Liu, W., Zhou, M., Wang, X., Qiu, W., Fan, L., Wang, B., & Chen, W. *The global burden of disease attributable to high fasting plasma glucose in 204 countries and territories, 1990-2019: An updated analysis for the Global Burden of Disease Study 2019.* Diabetes/metabolism research and reviews, 2022. **38**(8): e3572.
4. Göğüş, F. Z., Karlık, B., & Harman, G. *Classification of asthmatic breath sounds by using wavelet transforms and neural networks.* International Journal of Signal Processing Systems, 2014. **3**(2): p. 106-111.
5. Güler, İ., Polat, H., & Ergün, U. *Combining neural network and genetic algorithm for prediction of lung sounds.* Journal of Medical Systems, 2005. **29:** p. 217-231.
6. Yeginer, M., & Kahya, Y. P. *Feature extraction for pulmonary crackle representation via wavelet networks*. Computers in Biology and Medicine, 2009. **39**(8): p. 713–721.
7. Reichert, S., Gass, R., Brandt, C., & Andrès, E. *Analysis of respiratory sounds: State of the art.* Clinical Medicine: Circulatory, Respiratory and Pulmonary Medicine, 2008. p. 45-58.
8. Pasterkamp, H., & Zielinski, D. *The History and Physical Examination.* Kendig's Disorders of the Respiratory Tract in Children, 2019 (9th Edition). p. 2–25.
9. Sarkar, M., Madabhavi, I., Niranjan, N., & Dogra, M. *Auscultation of the respiratory system.* Annals of Thoracic Medicine, 2015. **10**(3): p. 158-168.
10. Zaitseva, E. G., Chernetsky, M. V., & Shevel, N. A. *About Possibility of Remote Diagnostics of the Respiratory System by Auscultation*. Devices and Methods of Measurements, 2020. **11**(2): p. 148-154.
11. Kim, Y., Hyon, Y., Jung, S. S., Lee, S., Yoo, G., Chung, C., & Ha, T. *Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning.* Scientific Reports, 2021. **11**(1): p. 1-11.
12. 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., & Lai, F. *Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database—hf_lung_v1*. PLOS ONE, 2021. **16**(7): p. 1-26.
13. Rocha, B. M., Filos, D., Mendes, L., Serbes, G., Ulukaya, S., Kahya, Y. P., … de Carvalho, P. *An open access database for the evaluation of respiratory sound classification algorithms.* Physiological Measurement. 2019. **40**: 035001
14. Jakovljević, N., & Lončar-Turukalo, T. *Hidden Markov Model Based Respiratory Sound Classification.* IFMBE Proceedings, 2017. **66**: p. 39–43.
15. Chambres, G., Hanna, P., & Desainte-Catherine, M. *Automatic Detection of Patient with Respiratory Diseases Using Lung Sound Analysis*. 2018 International Conference on Content-Based Multimedia Indexing (CBMI). 2018. p. 1-6.
16. Kochetov, K., Putin, E., Balashov, M., Filchenkov, A., & Shalyto, A. 2018. *Noise Masking Recurrent Neural Network for Respiratory Sound Classification.* In Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018. Springer International Publishing. p. 208–217.
17. Ma, Y., Xu, X., Yu, Q., Zhang, Y., Li, Y., Zhao, J., & Wang, G. *LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm.* 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS), 2019. IEEE. p. 1-4.
18. Ngo, Pham, L., Nguyen, A., Phan, B., Tran, K., & Nguyen, T. (2021). *Deep Learning Framework Applied For Predicting Anomaly of Respiratory Sounds.* 2021 International Symposium on Electrical and Electronics Engineering (ISEE). IEEE. p. 42-47.
19. Acharya, J., & Basu, A. *Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning.* IEEE Transactions on Biomedical Circuits and Systems, 2020. **14**(3): p. 535-544.
20. Serbes, G., Ulukaya, S., & Kahya, Y. P. *An Automated Lung Sound Preprocessing and Classification System Based OnSpectral Analysis Methods.* In Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017. Springer Singapore. p. 45-49.

21. Demir, F., Sengur, A., & Bajaj, V. *Convolutional neural networks based efficient approach for classification of lung diseases.* Health Information Science and Systems, 2019. **8**(1): 4.
22. Demir, F., Ismael, A. M., & Sengur, A. *Classification of lung sounds with CNN model using parallel pooling structure.* IEEE Access, 2020. **8**: p. 105376-105383.
23. ER, M. B. *Akciğer Seslerinin Derin öğrenme ile sınıflandırılması.* Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji, 2020. **8**(4): p. 830–844. (In Turkish).
24. Asatani, N., Kamiya, T., Mabu, S., & Kido, S. *Classification of respiratory sounds using improved convolutional recurrent neural network.* Computers & Electrical Engineering, 2021. **94**: 107367.
25. Fan, C.-Y., Liu, C.-P., Wang, K.-C., Jhan, J.-H., Wang, Y.-C. F., & Chen, J.-C. *Face Feature Recovery via Temporal Fusion for Person Search.* ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. 2020, IEEE. p. 1893-1897.
26. Liu, R., Cai, S., Zhang, K., & Hu, N. *Detection of Adventitious Respiratory Sounds based on Convolutional Neural Network.* 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS). 2019, IEEE. p. 298-303.
27. Perna, D., & Tagarelli, A. *Deep Auscultation: Predicting Respiratory Anomalies and Diseases via Recurrent Neural Networks.* 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). 2019, IEEE. p. 50-55.
28. Zulfiqar, R., Majeed, F., Irfan, R., Rauf, H. T., Benkhelifa, E., & Belkacem, A. N. *Abnormal respiratory sounds classification using deep CNN through Artificial Noise addition.* Frontiers in Medicine, 2021. **8**: 714811.
29. Nguyen, T., & Pernkopf, F. *Lung sound classification using co-tuning and stochastic normalization*. IEEE Transactions on Biomedical Engineering, 2022. **69**(9): p. 2872–2882.
30. Saraiva, A., Santos, D., Francisco, A., Sousa, J., Ferreira, N., Soares, S., & Valente, A. *Classification of respiratory sounds with convolutional neural network*. Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies. 2020, Science and Technology Publications. p. 138-144.
31. Ntalampiras, S., & Potamitis, I. *Automatic acoustic identification of respiratory diseases.* Evolving Systems, 2020. **12**(1): p. 69-77.
32. Krishnan, S. *Advanced Analysis of Biomedical Signals*. Biomedical Signal Analysis for Connected Healthcare, 2021: p. 157–222.
33. Li, L., Wu, Z., Xu, M., Meng, H. M., & Cai, L. *Combining CNN and BLSTM to Extract Textual and Acoustic Features for Recognizing Stances in Mandarin Ideological Debate Competition.* In Interspeech, 2016. p. 1392-1396.
34. Hakki, L., & Serbes, G**.** *Wheeze Events Detection Using Convolutional Recurrent Neural Network*. In 2023 Innovations in Intelligent Systems and Applications Conference (ASYU), Sivas, Turkiye, 2023. IEEE. p. 1-6.