Detecting Respiratory Diseases via Analysing Respiratory Sounds Across Diverse Population

Hemansh Shridhar, Ketan Shakhya, and Tushar Sandhan

Indian Institute of Technology (IIT) Kanpur
shridharhemansh@gmail.com,ketanshakya111@gmail.com,sandhan@iitk.ac.in

Abstract. In this paper, we present a machine learning-based approach for detecting respiratory problems using audio recordings of respiratory sounds captured from electronic stethoscopes and mobile devices. We utilised four distinct datasets to enhance the diversity and robustness of our training procedure, including the ICBHI dataset, which we expanded by adding COVID-19 and Hyperkinetic dysphonia class due to the critical importance of having diverse data in remote disease detection.Two datasets containing six and seven classes were created from these datasets and subsequently two different models were trained and tested on these datasets. Our methodology involved data augmentation techniques to improve the generalizability of our model. We employed two different models, Convolutional Neural Network - Gated Recurrent Unit Model and Bidirectional Long Short Term Memory Model, which are well-suited for processing sequential data like audio recordings. The models were trained and tested on these comprehensive dataset, achieving an overall score of 82.09% and 62.5% which are near to top performing approaches on similar datasets using similar metrics. This approach could significantly contribute to early diagnosis and monitoring, especially in areas with limited access to healthcare facilities.

Keywords: AI in Healthcare · Audio Signal Processing· Sequential Models.

1 Introduction

Because of their high frequency and potential severity, respiratory disorders such pneumonia, COVID-19, asthma, bronchitis, and chronic obstructive pulmonary disease (COPD) pose serious threats to world health [1]. The COVID-19 epidemic has brought to light the critical need for sophisticated diagnostic instruments that can deliver timely and reliable results. Existing techniques, such as stethoscope-based examination, frequently call for face-to-face patient engagement and are sensitive to the interpretation and experience of the physician [1], [2]. To overcome these limitations, we have explored the development of a machine learning-based model for remote detection of respiratory diseases. This model provides a scalable, yet effective way to remotely monitor respiratory health, with the goals of improving diagnostic accuracy, enabling early intervention, and supporting healthcare systems. An electronic stethoscope is a sophisticated tool that amplifies and records internal body sounds, offering enhanced clarity and detail compared to traditional stethoscopes. This precision makes it invaluable for capturing serious respiratory anomalies. Smartphones, on the other hand, are ubiquitous and equipped with high-quality microphones capable of recording audio in various environments. By leveraging the accessibility of smartphones alongside the precision of electronic stethoscopes, datasets encompassing a diverse range of respiratory sounds have recently been a major point of focus. Therefore training and testing machine learning models on such datasets ensures the generalizbility and robustness of such models, making them a comprehensive solution for remote respiratory disease detection [3], [4], [5], [6].

2 Related Works

Applications for speech and audio signals can be found in various fields, such as emotion recognition [7] and healthcare state recognition [8]. Audio analysis is just one of various essential procedures involved in the clinical diagnosis of voice related disorders. Finding disease-specific characteristics in respiratory sounds is essential for screening for breathing-related auditory disorders [9]. Pathologies ranging from inflammation and blockage to consolidation and pleural effusion are included in the category of respiratory diseases. While some traits may be shared by multiple diseases, pathology-related patterns like location and severity usually show variations particular to each disease [10].

In this section literature work related to detecting diseases involving audio data of patients across different patients has been discussed briefly. Some of the top performing and important approaches have been discussed in this section. Fagherazzi et al. [11] showcased a general pipeline for audio related disease detection. The general pipeline proposed int his work included a voice recorder to record the data, pre-processing the recorded audio, extracting relevant features from the audio, selecting suitable model for training and validation on extracted audio features and then subsequently integrating the algorithm on hardware edge devices such as smart phones etc for clinical practise.

Based on the workflow introduced in [11], Kim et al. [12] proposed a novel methodology named voice disorder detection using MFCC (Mel-frequency cepstral coefficients), fundamental frequency and spectral centroid was introduced . This study made use of both audio signals and patient data (gender, lifestyle etc) which included acoustical attributes such as spectral centeroid and fundamental frequency. Features were extracted in the form of MFCCs and BiLSTM Model was used for feature extraction and XGBoost classifier was used for classification. The patient data was analysed in the form of a dataframe, subsequently the dataframe was passed to an Artificial Neural Network. The ANN prediction output and the BiLSTM prediction outputs were stacked and the resultant feature was classified using an XGBoost classifier.The method showed remarkable results of accuracy of 95.67%, sensitivity of 95.36%, specificity of 96.49% and f1 score of 96.9%, outperforming existing techniques. As seen in the Introduction section the major Biomarker in any acoustic disease is presence and distribution of wheezes. Taking this into account detecting wheezes in the audio data is of paramount concern in detecting the presence or absence of acoustic disorders. The work introduced in [3] provided a standard method that took into account the standard patient data that is collected in a medical examination. In this work MFFCs extracted were fed to ResNet-34, the tabular data that contains the patient information was also utilised. The audio features and patient data was simultaneously fed to fully connected layers. The features were concatenated and finally fed to a classifier that classifies between presence and absence of wheezes.

Some of the major works that deal with the segmentation and detection of adventitious segments are discussed here. ICBHI is the largest publicly available dataset of respiration sounds. This dataset contains the segment-wise information of the presence and absence of wheezes and crackles, along with the disease label. In [13], the authors utilized Audio Spectrogram transformer architecture pre -trained on large-scale visual and audio data to generalise well on the respiration classification task. This work introduced a straightforward Patch-Mix augmentation, which randomly mixes patches between different samples. Along with this a Patch-Mix contrastive loss was applied to distinguish between the mixed representations. Another major problem area in the field of respiratory disease classification is the addition of device dependent features in the audio data which shows up in the frequency representations. Nguyen et al. [6] showed a method to remove the frequency response of the device used during the collection of data samples using stochastic normalisation. Thus the domain adaptation based approaches that remove the inter class similarity present between different class samples is another major problem area. Kim et al. [14] introduced device guided contrastive learning approach that effectively reduces device-induced similarity within the data. Metadata collected using clinical examination can act as additional information that can effectively improve the metrics. In [15] utilised supervised contrastive learning using metadata class labels to learn useful representations, and proposed an extension with multiple pretext tasks for a performance boost. This work showed that combining cross entropy loss with supervised contrastive loss can be used to improve the classification metrics.

3 Dataset Description

The major requirement of any Voice Disorder detection system that is effective in real life application is the ability to generalise well over diversely recorded data as well as data recorded through different medical examination techniques Therefore the need of a dataset containing samples recorded using different recording devices as well as different examination technique becomes imminent. To simulate such conditions combined audio samples from 4 publicly available datasets were combined. The datasets utilised were recorded using different electronic devices with same class having samples recorded using different electronic devices. In this work the ICBHI dataset [16], which includes respiratory sounds

Table 1: Table Showing the recording Devices as well as disease class in each dataset

Dataset	Classes taken from the dataset	Electronic Instruments used	Sampling Frequency
VOICE dataset	Hyperkinetic Dysphonia	Smartphone Microphone (Samsung Galaxy)	8000 Hz
Respiratory Sound Dataset	Asthama , Bronchitis , COPD ,Healthy ,Pneumonia	AKG C417L Microphone (AKGC417L), Littman Classic SE Stethoscope, Electronic Stethoscope, Master Elite Electronic Stethoscope	44100 Hz
Electronic Stethoscope Sound	Bronchitis , COPD , Healthy ,Pneumonia	Electronic Stethoscope	4000
COSOWARA	COVID-19	Smartphone microphone	48000 Hz

captured by electronic stethoscopes and smartphone microphones has been used along with datasets such as, “the Electronic Stethoscope Dataset of Lung Sounds Recorded from the Chest Wall” [17], which contains respiration sounds of people with diseases like asthma, COPD, lung fibrosis, heart failure, and pneumonia. This added more data to the classes of ICBHI dataset which had very low data for classes like Asthma, and with the help of two more datasets, COUGHVID [18] and VOICED [19] two more classes of disease COVID-19 and Hyperkinetic Dysphonia were included. The summary of the details including the device information of various datasets has been included in table 1. From the table, the recording diversity of the datasets is easily observable along with different sampling rate across data-points of the same data class.

As shown in Table 1 , The ICBHI [16] dataset was constructed using four distinct types of stethoscopes: Meditron, LittC2SE, Litt3200, and AKGC417L, each containing a unique input sensor. The Litt3200 utilizes a piezoelectric sensor to capture low frequencies, such as heart murmurs. The AKGC417L and Meditron stethoscopes employ condenser microphones, which are optimal for high frequencies like wheezes and provide superior resolution across a broader frequency range. The LittC2SE is an analog stethoscope featuring a dynamic microphone, making it ideal for mid frequencies and offering a balanced performance for general auscultation. In VOICED dataset contains audio signals consisting of a clinical examinations done by recording of vocalization of the vowel ’a’ by the patient five seconds in length without any sound interruption. The addition of this data was to include the variation in examination method in the dataset. This is done with view that performance of any respiratory disease detection system should be independent of the method of examination. COUGHVID dataset has been used to add the Covid-19 class to the data . Since cough is one of the major symptoms of Covid-19, this class has been added to the dataset to include the important cough feature, which is one of the most common indicator of respiration related disorder.This dataset contains 25,000 crowd sourced cough recordings. This dataset contains 2,800 labeled recordings. 50 samples belonging to different individuals were taken from this dataset to represent the Covid-19 class. Electronic Stethoscope sounds was used to include various classes namely Bronchitis, COPD ,Healthy ,Pneumonia. This dataset was collected using electronic stethoscope which has been used to record lung sounds from healthy and unhealthy subjects. This dataset contains seven disease classes namely asthama, heart failure, pneumonia, bronchitis, pleural effusion, lung fibrosis and COPD as well as healthy breathing sounds. The lung sound was recorded by placing stethoscope at different chest locations. The stethoscope placement was determined by a specialist physician.

4 Experimentation

4.1 Evaluation Metrics

We employed Sensitivity (Se), Specificity (Sp), and a combined Score metric as our performance indicators for our dataset as used in ICBHI[16] challenge. Sensitivity, or the true positive rate, measures the proportion of actual disease cases correctly identified by the model. Specificity, conversely, measures the ratio of healthy individuals accurately classified as disease-free to the total number of samples for the healthy class. To strike a balance between these two metrics, the Score metric has been used, which is calculated as the arithmetic mean of Sensitivity and Specificity.

4.2 Pre-processing and feature extraction

The audio recordings present in all the four datasets that have been used to create the 6 class and 7 class dataset, are sampled with a rate that varies from 8 kHz to 48 kHz. All the audio files were resampled to 16 kHz mono as done in previous studies [20]. The length of each audio sample has been limited to 10 seconds. In case of the VOICED dataset where the length of the sample is less then 8 seconds , the audio samples are padded with zero values.

In this work we perform different feature extraction procedure on the six class dataset and the 7 class dataset. The effectiveness of data augmentation in respiratory sound classification has been proved in [21] , [22]. Hence, we perform data augmentation on the raw audio files. A 80-20% random split was done on the data to obtain the training and testing samples. For each audio sample in both the dataset we adopt three new augmentation strategies of noise addition, time shifting and stretching. This was done with the sole view of increasing the size of the dataset.The scaling factor for noise addition has been set at 0.001, audio stretching factor is set at 1.2 and the time shifting parameter is set at 1600. All these augmentations were done on the training data. From the resulting audio data MFCCs were extracted. For each audio sample 52 Mel Frequency Cepstral Coefficients were extracted. These features serve as input to the GRU-CNN architecture. For the second experiment similar augmentation strategies were applied on raw audio to increase the number of samples but instead 13 LFCCs were extracted for each audio sample. To get the robustness of the features extracted the extracted where embedded onto two-dimensional space to observe the amount of clustering among similar data samples.The distribution of class labels after augmentations is shown in Fig. 1 and Fig. 2

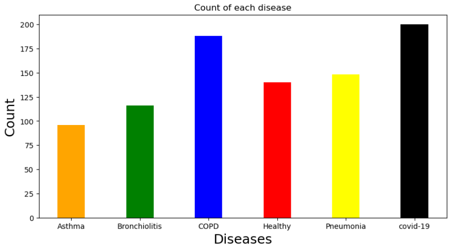

Fig. 1: The distribution of the six class dataset after augmentation is shown in the figure

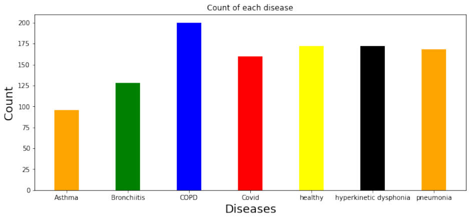

Fig. 2: The distribution of the seven class dataset after augmentation is shown in the figure

4.3 Model Architecture

CNN-GRU Model The CNN-GRU architecture contains two blocks each containing 1- Dimensional convolution block, Maxpooling followed by a Batch-normalisation layer at the bottom of the architecture. The convolutional layer of the first and second block consists of 256 and 512 filters each. Each convolution layer is followed by a Max pooling layer and a Batch Normalisation layer. From the input containing 52 features this block extracts 512 features, these extracted temporal features are subsequently fed to a network of GRUs aimed at capturing the temporal information from the extracted features.

The subsequent network of GRUs contains three branches each containing two GRU layers. The top layer of first and third branch consists of 32 units each while the second branch has a GRU layer with 64 units. Second layer of each branch consists of 128 units. Here, the resulting input shape has features equal to the number of units in the GRU layer. All the outputs of the branches are added element-wise and the resulting output has 128 sequence coefficients. The output at this stage is subsequently passed to another GRU network consisting of two branches in a similar fashion containing two layers each. The top layers contain 32 and 64 GRU units respectively whereas the second layers contain 32 units each. A Residual like connection is between the first network and the second network. The output of the GRU layer in the top layer of first network containing 32 units is added element wise to the output of the second network. The output obtained at this stage is fed to a containing two branches of dense fully connected layers. The first branch consists of two layers mapping the output to 32 and 128 feature coefficients respectively. The second branch of the network maps the output obtained to 64 and 128 features respectively. The resultant output obtained containing 128 features from both the branches is added element-wise and passed to a final fully connected classifier mapping the output to 6 class scores. The architecture of the CNN-GRU model is shown in Fig. 3

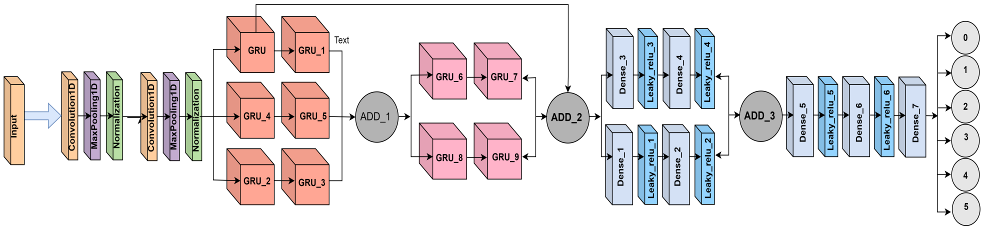

Fig. 3: The figure shows the architecture of the CNN-GRU model

Bidirectional Long Short Term Memory Bidirectional Long Short-Term Memory (BiLSTM)[23] networks are a powerful variant of recurrent neural networks that have gained significant popularity in various research projects, particularly in natural language processing, speech recognition, and time series analysis. BiLSTMs process input sequences in both forward and backward directions, allowing the network to capture context from both past and future states. Thus BiLSTM present them as a powerful tool to analyse data that has temporal information contained inside them like audio speech signals. The second experimentation used an architecture that contained 16 BiLSTM layers stacked on top of each other followed by a fully connected layer that acts as classifier classifying the output into 7 disease classes. In this model architecture, 16 BiLSTM layers were stacked on top of each other. Each BiLSTM layer generates an output sequence, this output sequence is subsequently passed to the next layers.The concatenated hidden states from the last layer of the BiLSTM layers are passed onto the classifier which calculates the score for each class.The architecture of this model is shown in Fig. 4

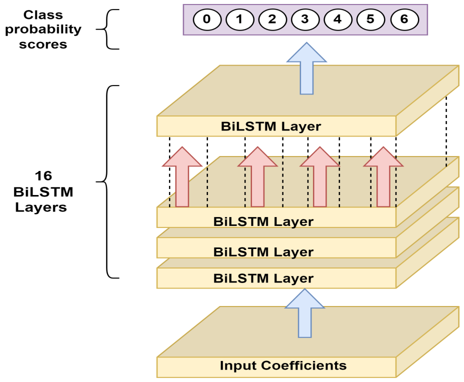

Fig. 4: The figure shows the architecture of the BiLSTM model

4.4 Methodology

Experiment 1 In experiment 1 we have used a dataset containing six classes namely Asthama, Bronchiolitis, COPD, Healthy, Pneumonia and Covid-19 obtained from 3 different datasets. The main aim of the experiment is to test the model performance on data emulating diverse clinical and recording conditions. Since the dataset contains samples recorded using different electronic recording devices and recorded from different body location, we assume it successfully serves this purpose. For this dataset we have used the CNN-GRU model where the GRUs are stacked on top of two blocks containing CNN, Max-pooling and Batchnormalisation layers. In the CNN-GRU architecture both CNN and GRU are used to leverage the strengths of both kind of layers in capturing both the spatial and temporal feaures of the data. CNNs are adept at detecting local patterns in the data, such as edges in images or short-term features in audio signals. By applying convolutional filters Since MFCCs and LFFCs contain information such as amplitude value corresponding to different frequency components, CNNs learn hierarchical representations present within the MFCCs, capturing low-level to high-level features. Moreover valid Biomarker such as wheezes and crackles are present locally within certain time instants. CNNs are effective in capturing such local features. CNNs are used here to extract more useful features from the given input. The bottom two blocks containing the CNN layers along with the Max Pooling and the Batch-normalisation project the given 52 feature coefficients to 512 coefficients. The features extracted by CNN are passed onto a network of GRU where each GRU computed the input sequential feature coefficients in a reverse manner. The equations for a Gated Recurrent Unit (GRU) are defined as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{1}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{2}$$

$$\tilde{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1}) + b_h) \tag{3}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{4}$$

where:

- z_t is the update gate.
- r_t is the reset gate.
- $\tilde{h}_t$ is the candidate activation.
- h_t is the hidden state at time step t.
- x_t is the input at time step t.
- σ is the sigmoid activation function.
- tanh is the hyperbolic tangent activation function.
- $W_z, W_r, W_h, U_z, U_r, U_h$ are weight matrices.
- b_z, b_r, b_h are bias vectors.
- $\odot$ denotes element-wise multiplication.

The GRU network contains two blocks containing three and two branches where at the end of each block the output of each branch is added element wise. A residual connection has been made from the first block contінaining GRU branches to the second block . The residual connection enables the output of the second block to incorporate information from the first block, leading to more comprehensive feature representations. This has been done with purpose improving the model's robustness to variations in the input data and enhance its ability to generalize to test data. The residual connection allows the model to reuse features learned in the earlier block. Since audio information projected to time frequency domain is generally abstract in nature , adding residual connections allow the model to reuse features learned in the earlier block. Finally a classifier layer was added on top of dense fully connected layers to classify the extracted features to one of the six classes.

Experiment 2 In the second experiment a new dataset containing seven classes, six classes from the dataset used in experiment 1 has been used along with hyperkinetic dysphonia being the newly introduced class. For this datasets we have extracted LFCCs as features to be passed to the subsequent model. We have used different feature extraction technique to test the performance of this feature extraction technique. 13 LFCCs for each audio file were extracted from the audio data in the dataset. The model containing stacked BiLSTM has been trained and tested on this dataset. Stacking multiple BiLSTM layers allows the model to learn increasingly abstract representations of the input sequence. Each layer captures temporal dependencies at different levels of granularity, enabling the model to understand complex temporal relationships in the data. Each layer's forward and backward hidden states are concatenated and passed to the next layer as sequences. The output of last layers is averaged and passed onto the fully connected layers outputting the probability scores of the 7 classes.

Table 2: Comprehensive comparison of different methods for the respiratory sound classification task.

Methods	SP%	SE%	AS%	Params(in M)	No of Classes
LungAttn[24]	71.44	36.36	53.9	0.7 (estimated)	4
Co-Tuning[6]	79.34	37.2	56.2	21(estimated)	4
RespireNet[3]	72.30	40.10	56.2	4.3(estimated)	4
SG-SCL[14]	79.87	43.55	61.71	-	4
AST+Patch-Mix CL[13]	81.66	43.07	62.37	21(estimated)	4
BiLSTM(ours)	62.5	80	71.25	6(approx)	7
CNN-GRU	64.28	99.9	82.09	12(approx)	6

5 Results

The results of Sensitivity (Se), Specificity (Sp), and Score obtained from both the models on the 6 class dataset and the 7 class dataset have been summarised in Table .

Since there is absence of any work related to composite dataset obtained by mixing publicly available datasets. Thus rather we compare the results with Baseline obtained on ICBHI Respiratory Sound Database .Since it is one of the constituent components of our datasets . On 6 class dataset we have obtained a value of 64.28% for specificity , 99.9% value for sensitivity and overall score of 82.09%. On 7 class dataset using BiLSTM model we calculated values of 62.5% for specificity , 80% for sensitivity and overall score of 71.25% . The average score obtained is well above the standard works on ICBHI dataset as shown in Table 2. Moreover on further analysis we can observe that the number of trainable parameters are less as compared to most of the standard works on ICBHI dataset, dealing with respiration abnormality detection. The confusion matrix for both the models have been shown in Fig. 5 and Fig. 6

6 Conclusion

In this work we have trained and tested models on datasets containing data from four different that perfectly capture the variation across different domains namely recording device, patient age, patient ethnicity and examination method. The proposed models proved their effectiveness by achieving scores of 82.09% and 71.25%. Out of which CNN-GRU showed remarkable score value proving its effectiveness in being a suitable candidate for detecting respiratory anomalies on real world data. The uniqueness of this architecture stems from the fact that it contains CNN layers to capture local spatial features as well as GRU blocks that are efficient in capturing the temporal features of the data. Despite the presence of various common domain related similarities between the data points of various classes, the proposed approach has shown remarkable results.

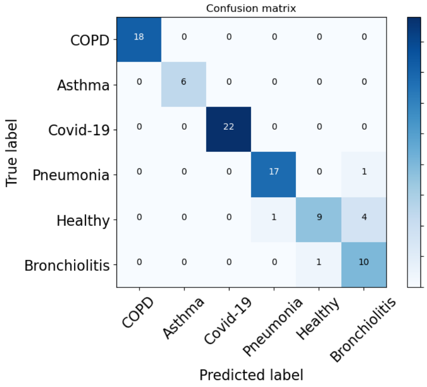

Fig. 5: The figure shows the confusion matrix obtained on the 6 class dataset using the CNN-GRU model

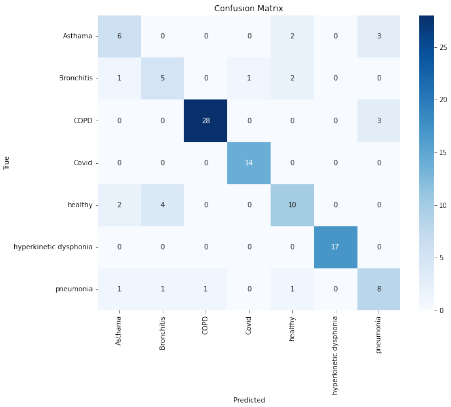

Fig. 6: The figure shows the confusion matrix obtained on the 7 class dataset using the BiLSTM model

References

1. M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, “Auscultation of the respiratory system,” *Annals of thoracic medicine*, vol. 10, no. 3, pp. 158–168, 2015.
2. L. Arts, E. H. T. Lim, P. M. van de Ven, L. Heunks, and P. R. Tuinman, “The diagnostic accuracy of lung auscultation in adult patients with acute pulmonary pathologies: a meta-analysis,” *Scientific reports*, vol. 10, no. 1, p. 7347, 2020.
3. S. Gairola, F. Tom, N. Kwatra, and M. Jain, “Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 527–530.
4. Z. Ren, T. T. Nguyen, and W. Nejdl, “Prototype learning for interpretable respiratory sound analysis,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9087–9091.
5. Z. Wang and Z. Wang, “A domain transfer based data augmentation method for automated respiratory classification,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9017–9021.
6. T. Nguyen and F. Pernkopf, “Lung sound classification using co-tuning and stochastic normalization,” *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 9, pp. 2872–2882, 2022.
7. M. S. Hossain, G. Muhammad, M. F. Alhamid, B. Song, and K. Al-Mutib, “Audio-visual emotion recognition using big data towards 5g,” *Mobile Networks and Applications*, vol. 21, pp. 753–763, 2016.
8. M. S. Hossain, “Patient state recognition system for healthcare using speech and facial expressions,” *Journal of medical systems*, vol. 40, pp. 1–8, 2016.
9. B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni *et al.*, “An open access database for the evaluation of respiratory sound classification algorithms,” *Physiological measurement*, vol. 40, no. 3, p. 035001, 2019.

10. W.-c. Dai, H.-w. Zhang, J. Yu, H.-j. Xu, H. Chen, S.-p. Luo, H. Zhang, L.-h. Liang, X.-l. Wu, Y. Lei *et al.*, "Ct imaging and differential diagnosis of covid-19," *Canadian Association of Radiologists Journal*, vol. 71, no. 2, pp. 195–200, 2020.
11. G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for health: the use of vocal biomarkers from research to clinical practice," *Digital biomarkers*, vol. 5, no. 1, pp. 78–88, 2021.
12. B. J. Kim, B. S. Kim, J. H. Mun, C. Lim, and K. Kim, "An accurate deep learning model for wheezing in children using real world data," *Scientific Reports*, vol. 12, no. 1, p. 22465, 2022.
13. S. Bae, J.-W. Kim, W.-Y. Cho, H. Baek, S. Son, B. Lee, C. Ha, K. Tae, S. Kim, and S.-Y. Yun, "Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification," *arXiv preprint arXiv:2305.14032*, 2023.
14. J.-W. Kim, S. Bae, W.-Y. Cho, B. Lee, and H.-Y. Jung, "Stethoscope-guided supervised contrastive learning for cross-domain adaptation on respiratory sound classification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2024, pp. 1431–1435.
15. I. Moummad and N. Farrugia, "Pretraining respiratory sound representations using metadata and contrastive learning," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* IEEE, 2023, pp. 1–5.
16. B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques *et al.*, "A respiratory sound database for the development of automated classification," in *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017.* Springer, 2018, pp. 33–37.
17. M. Fraiwan, L. Fraiwan, B. Khassawneh, and A. Ibnian, "A dataset of lung sounds recorded from the chest wall using an electronic stethoscope," *Data in Brief*, vol. 35, p. 106913, 2021.
18. L. Orlandic, T. Teijeiro, and D. Atienza, "The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, p. 156, 2021.
19. U. Cesari, G. De Pietro, E. Marciano, C. Niri, G. Sannino, and L. Verde, "A new database of healthy and pathological voices," *Computers & Electrical Engineering*, vol. 68, pp. 310–321, 2018.
20. W. Song, J. Han, and H. Song, "Contrastive embeddind learning method for respiratory sound classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2021, pp. 1275–1279.
21. J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *IEEE transactions on biomedical circuits and systems*, vol. 14, no. 3, pp. 535–544, 2020.
22. J. T. C. Ming, N. M. Noor, O. M. Rijal, R. M. Kassim, and A. Yunus, "Advanced and minor lung disease severity classification using deep features," in *2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA).* IEEE, 2019, pp. 122–127.
23. A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.
24. J. Li, J. Yuan, H. Wang, S. Liu, Q. Guo, Y. Ma, Y. Li, L. Zhao, and G. Wang, "Lungattn: advanced lung sound classification using attention mechanism with dual tqwt and triple stft spectrogram," *Physiological Measurement*, vol. 42, no. 10, p. 105006, 2021.