Fusion of Manual and Deep Learning Analyses for Automatic Lung Respiratory Sounds Identification in Youth

Behrad TaghiBeyglou[1,2], Atousa Assadi[1,2], Ahmed Elwali[1,2] and Azadeh Yadollahi[1,2]

[1] Institute of Biomedical Engineering, University of Toronto, ON, Canada
[2] KITE Research Institute, Toronto Rehabilitation Institute – University Health Network, ON, Canada

Abstract— **Lung sounds contain important clinical information which can be used for identifying respiratory and/or lung disorders. Manual identification of respiratory events is time-consuming and prone to subjective errors. While several automatic respiratory event classification techniques have been proposed previously, they are mostly focused on the identification of respiratory sounds in the adult population. Though, this is challenging in youth as lung is developing till the age of 20 years old which affects the parameters of respiratory sounds. In this research, our goal is to develop techniques for respiratory sound classification in youth using the SPRSound dataset, which includes recordings of individuals from 0-18 years old. The objectives include binary and multi-class classification of respiratory events (objective 1) and recordings (objective 2). For objective 1, we extracted purified respiratory features using a convolutional neural network (CNN) as well as frequency and time domain features, statistical features, and patient demographics, while for objective 2, a mixed model of long short-term memory (LSTM) network and a gradient boosting classifier with a novel voting scheme is developed. The features which were significantly associated with the different respiratory sounds were used to train machine-learning models for classification purposes. We evaluated the models' performance based on sensitivity, specificity, an average of sensitivity and specificity scores (AS), and the F1-score. The final performance score is defined as the average of the AS and F1-score. Our proposed framework reached 0.91±0.03 and 0.82±0.03 in binary and 7-class event classification, respectively. Also, the developed model reached 0.74±0.02 and 0.55±0.03 in 3-class and 5-class recording sound classification.**

Keywords— **Respiratory Sounds, Model Fusion, Deep Learning, Signal Processing, Paediatric.**

I. Introduction

Despite having many superior lung imaging diagnosing techniques, lung auscultation is still the most common, accessible, and affordable way for initial examination of individual's respiratory sounds for diagnosis and detection of abnormal breathings and their associated disorders [1, 2]. Abnormal breathing sounds include a mixture of normal and adventitious breathing events such as wheeze, rhonchus, stridor, fine and coarse crackles. Respiratory event sounds contain distinctive features in the temporal (happens at a certain breathing phase and/or for a specific duration of time) and frequency (happens in particular frequency ranges) domains, which can be used to distinguish normal vs. adventitious events: Wheeze is characterized by musical and high-pitched sounds with fundamental frequency of >500 Hz; Rhonchus is a variant of the wheeze with lower pitch, typically near 150 Hz; Stridor is a high-pitched, musical sound with the fundamental frequency of approximately 500Hz; Fine crackles are brief, discontinuous, popping high-pitched lung sounds with typical frequency of 650Hz and duration of 5ms; Coarse crackles are discontinuous, brief, popping lung sounds with typical frequency of 350Hz and duration of 15ms.

With the advancement of technology and data analysis techniques, breathing sounds can be recorded reliably (with digital stethoscope) and automatically analyzed for identifying breathing disorders using machine learning and artificial intelligence techniques. For example, automatic breathing sound analysis have been used for diagnosing sleep apnea, asthma, and COPD [1–3]. For instance, some teams used tracheal breathing sounds to automatically diagnose obstructive sleep apnea during sleep or screen it during wakefulness [4].

While automatic identification of breathing sounds can improve diagnosis, previous studies are mostly focused on adult patients [5–7]. Automatic identification of breathing sounds in youth is challenging mainly because human respiratory system is still developing [8]. Human lung gets mature around the age of 20 years old. The number of lung air sacs increases rapidly within the first three years of individuals' lives; then the lung becomes the same as the adult's lung except in size [9, 10]. From 3 years old to 10 years old, the size of the lung increases with height. Moreover, sex, age, height, weight, and puberty have different impacts on the size of the lung and lung function and accordingly breathing sounds. For instance, between the ages of 3-10 years old, females show greater forced airflows per unit of lung volume than males [9, 11]. Therefore, these confounding factors should be considered while analyzing breathing sounds for diagnosis in individuals between 0 – 18 years old.

As a result, in this research, we aim to develop methods for automatic classification of sound recordings and breathing sounds in the young population. To achieve this goal, we aim to address two objectives:

1. To develop a technique for automatic identification of breathing events in the following settings
   - Binary classification: normal vs. adventitious
   - Multiclass classification: normal, rhonchus, wheeze, stridor, and coarse crackles, fine crackles, wheeze & crackles
2. To develop a technique for automatic labelling of signal recordings in the following settings:
   - Ternary classification: normal, adventitious, poor quality
   - Multiclass classification: normal, continuous adventitious sound (CAS), discrete adventitious sound (DAS), mixture of CAS and DAS, or poor quality

II. Methods

A. Dataset

The dataset used in this research was initially developed for BioCAS 2022 Grand Challenge on Respiratory Sound Classification [12] and it was later extended and published as SPRSound dataset [13, 14]. Lung sounds are collected at Shanghai Children's

Medical Center from 251 participants (123 females) ranging from 1 month to 18 years old. Each participant has multiple recordings from different recording locations, such as left posterior, left lateral, right posterior, and right lateral. Each recording is accompanied by an annotation file which provides information regarding the label and duration of recording and respiratory events that occur during recording. Recording labels indicate the status of the whole recording, such as normal, CAS, DAS, CAS and DAS, and poor quality, while event labels reflect respiratory events (normal, rhonchus, wheeze, stridor, coarse crackle, fine crackle, and wheeze and crackle). The total number of recordings is 1949 and recordings are converted to digital .wav files using 8000Hz sampling rate.

B. Pre-processing

Since lung auscultation recordings are low-frequency content, they can interfere with cardiac sound vibrations [15]. Therefore, in this research, a 6th-order Butterworth filter with two different passbands (i.e., 75-850Hz and 75-1500Hz) was applied to raw recordings. The frequency range of 75-850Hz was used for the event detection task, while the frequency range of 75-1500Hz was used to capture parts of unwanted sources to make the model capable of classifying poor-quality recordings [16]. To avoid undesirable complications due to the nonlinear phase of the filter, the forward and backward method was used to keep the filter zero phase.

C. Objective 1: Respiratory event classification

To address objective 1, we first analyzed breathing events to extract the following event-related features: 1) time-frequency purified features using a 4-layer binary CNN model, which is trained on Mel spectrograms of breathing events, 2) statistical features, e.g. mean, variance, skewness, and kurtosis of signal, 3) nonlinear features, e.g. Shannon entropy and histogram logarithm, 4) spectral features, e.g. power of different frequency ranges (70-150Hz, 150-300Hz, 300-400Hz, 400-500Hz, 500-600Hz, 600-700Hz, 700-850Hz), 13 Mel frequency cepstral coefficients (MFCC) [17], and its 1st and 2nd derivatives, 5) temporal features, such as zero-crossing rate, 6) tempogram-related features such as autocorrelation, 7) demographic-related features such as age, sex, and 8) recording related features such as the location of recordings. The CNN model was trained as a classifier using a batch size of 32 samples and the Adam optimizer with a learning rate of 1E-4.

The total number of extracted features was 109. The extracted features were then fed to a feature selection pipeline (Python Featurewiz [18]) to remove zero variance and highly correlated features, leaving only those features which were strongly correlated with breathing event types. The selected features were then fed to binary (Sub-Objective 1.1) and multiple (Sub-Objective 1.2) classifiers including extreme gradient boosting (XGBoost), Random Forest, Logistic Regression, and Gradient Boosting models for comparing the classification performance and selecting the optimal classification model. The pipeline is shown in Fig. 1. To account for a highly imbalanced dataset, we used a dynamic loss function for our CNN model in which the weight of each class is computed dynamically based on the number of samples per class and the class with lower samples is penalized more for misclassification. Moreover, we trained the Featurewiz based on a balanced dataset in which we randomly selected the same number of normal vs. adventitious breathing events.

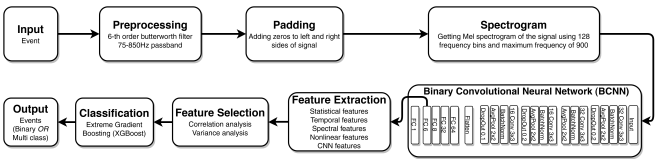


Fig. 1: Proposed pipeline for event classification (Zoom-in for more details).

D. Objective 2: Recording classification

To address objective 2, a new model based on long short-term memory (LSTM) neural network was developed. The training specifications for the proposed LSTM are similar to those of the CNN model in objective 1. In this objective, the inputs are recording samples (instead of event samples) and the outputs (labels) are recording labels. Recording labels can be either a three-state label (normal, adventitious, poor quality for Sub-Objective 2.1) or five-class output (normal, CAS, DAS, CAS and DAS, poor quality for Sub-Objective 2.2). As shown in Fig. 2, after zero-padding of the recordings at the beginning and end of the signal, the recording is segmented using a rectangular window with a length of 50ms and an overlap of 25ms to capture the temporal variation of the shortest event in the recording. From each segment, a set of features similar to those of objective 1 are extracted. In this objective, the overall model is based on two submodels. The first submodel uses LSTM to classify the recordings and generate each class probability P_{1i} (i can be either 1,2,3 or 1,2,3,4,5 depending on the number of output classes). The second submodel merges all the segments using the following equation. $\mu(f_i)$ demonstrates the average of feature f_i through all the segments of the recording and $\sigma(f_i)$ corresponds to the standard deviation of feature f_i over all segments of a recording.

$$f_i = \frac{\mu(f_i)}{\sigma(f_i)} \tag{1}$$

After merging the features, correlation and variance analyses (Featurewiz) are used to select the best set of features. An XGBoost model is then used to extract the class probabilities P_{2i} (i can be either 1,2,3 or 1,2,3,4,5). A voting scheme is developed for fusing two models using different (coefficients of the first submodel) and (coefficients of the second submodel). α_i and β_i values are ranged from 0 to 10 and calculated using a single-parameter grid search based on the results of classifying the validation set. The coefficients are then averaged throughout different runs to increase the model generalizability. Finally, the class with the highest score is selected as the result of classification (either ternary or five-class).

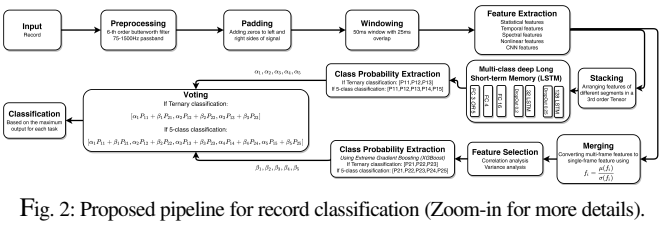


Fig. 2: Proposed pipeline for record classification (Zoom-in for more details).

E. Model Evaluation

We trained, validated, and tested our classification models using 70%, 15%, and 15% of the data, respectively. We evaluated the performance of our models based on the average and standard deviation of the overall score and performance measures using a 5-fold cross-validation technique in objective 1 and over 5 different runs in objective 2. The performance measures and overall score of the model are defined as:

- $SP = \frac{(\#\text{correctly predicted adventitious events/records})}{(\#\text{total adventitious events/records})}$
- $SE = \frac{(\#\text{correctly predicted normal events/records})}{(\#\text{total normal events/records})}$
- $AS = \frac{SE+SP}{2}$
- $HS = \frac{2*SE*SP}{SE+SP}$
- $Score = \frac{AS+HS}{2}$

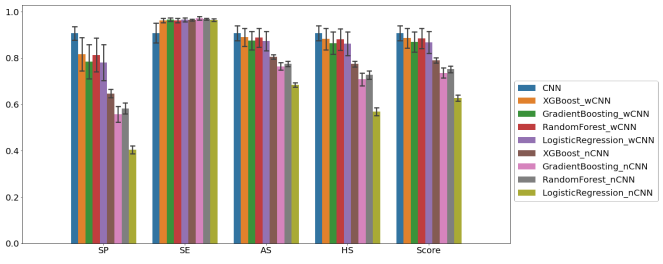
III. RESULTS

A. Binary Event Classification (Sub-Objective 1.1)

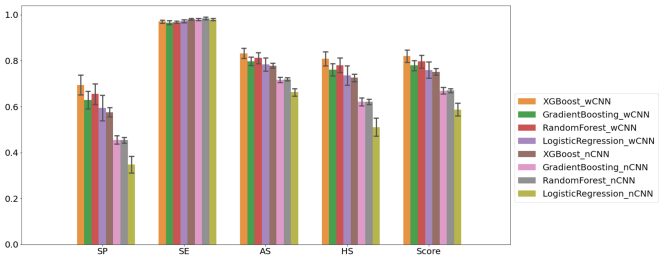
The results of the binary event classification models are presented in Table 1 and Fig. 3-a. The results indicated that CNN had the highest overall score. Though, the score was not significantly different from XGBoost, gradient boosting, logistic regression and random forest classification models which include both CNN and manually extracted features. Moreover, our results showed that the classical models infused with CNN features performed better in classifying adventitious events. Finally, the features extracted from CNN noticeably improved the performance of classical machine learning models.

B. Multiple Event Classification (Sub-Objective 1.2)

The results of the multiple event classification models are presented in Table 1 and Fig. 3-b. The results indicated that XGBoost infused with CNN features had the highest overall score. Though, the score was not considerably different from gradient boosting, logistic regression and random forest classification models with infused CNN features. Moreover, our results showed that extracted features from CNN significantly increased the performance of our machine learning models. Furthermore, based on our results, the developed classifiers performed better in classifying non-normal events than normal events. This is mainly because the Mel spectrum of the non-normal events are more localized due to their specific frequency ranges and physiological characteristics.



(a) Binary



(b) 7-class

Fig. 3: Performance of proposed model for objective 1, where **w** and **n** indicate the model **with** and **without** the CNN features, respectively.

C. Ternary Record Classification (Sub-Objective 2.1)

Using the min-max feature normalization technique (which maps the minimum to 0 and maximum to 1) for the second submodel, the results of ternary classification are acquired and shown in Table 1. To overcome the challenge of an imbalanced dataset, the loss function is normalized using class weights which are reciprocal to the number of samples in each class.

D. Multiple Record Classification (Sub-Objective 2.2)

For this task, the min-max feature normalization technique was used to encounter the challenge of different scales and units among the features. The results of this subtask are shown in and Table 1. It can be noted that the accuracy of this model is relatively high for a 5-class classification task. However, the sensitivity is low which means that the model was not capable of detecting adventitious recordings well. This phenomenon happened mainly because of the extremely low number of samples in each class.

As can be seen from the figure, the 5-class classification generally resulted in poorer sensitivity or SE in comparison to the ternary model which is due to a high imbalanced dataset that could not be completely addressed even if in presence of the weighted classification technique.

E. Overall performance

To align with the BioCAS Grand Challenge organizers scoring policy the following metric is used to estimate the overall performance of the proposed model.

$$\text{Total Score} = 0.2*\text{Score}_{1-1} + 0.3*\text{Score}_{1-2} + 0.2*\text{Score}_{2-1} + 0.3*\text{Score}_{2-2} \quad (2)$$

Table 1: Overall performance of the developed framework

Sub-Objective	Metrics				
	SP	SE	AS	HS	Score
1.1	0.91±0.03	0.91±0.04	0.91±0.03	0.91±0.03	0.91±0.03
1.2	0.69±0.04	0.97±0.01	0.83±0.02	0.81±0.03	0.82±0.03
2.1	0.78±0.03	0.69±0.02	0.74±0.02	0.74±0.02	0.74±0.02
2.2	0.76±0.07	0.4±0.04	0.58±0.03	0.52±0.03	0.55±0.03

Table 2: Comparison of the model with the top-5 teams

Method	Metrics				
	Score_{1-1}	Score_{1-2}	Score_{2-1}	Score_{2-2}	Total Score
1st Team [19]	0.89	0.82	0.72	0.53	0.73
2nd Team [20]	0.82	0.74	0.71	0.53	0.67
3rd Team [21]	0.85	0.75	0.70	0.53	0.69
4th Team [22]	0.90	0.80	0.72	0.45	0.70
5th Team [23]	0.84	0.73	0.67	0.52	0.68
Proposed model	**0.91**	**0.82**	**0.74**	**0.55**	**0.74**

[1] The original ranking included runtime as a bonus point in the total score (which is ignored in this table) [12].

Using the average of each subtask's performance to calculate the final score, the total score of 0.7387 is achievable by applying the proposed framework of this research. Also, the performance of our proposed model is compared with the top-5 teams of the competition in Table 2, and as it can be conceivable the suggested model outperformed the others.

IV. Discussion

Our results indicated that features extracted from the CNN model significantly improved the performance of our classical machine learning techniques. This is mainly due to the fact that deep learning models can introduce data-driven features which might be missed when only using manual feature extraction.

Additionally, in the second objective, using a new voting scheme to merge the deep LSTM model with XGBoost helped us achieve better and more robust results in comparison to using each one separately. The fusion was applied mainly because the submodels performed contrastingly. For example, the LSTM-based submodel focused on normal and poor quality recordings, while the XGBoost-based was aimed to extract and characterize adventitious recordings.

Furthermore, the problem of imbalanced data is the inevitable state-of-the-art challenge of artificial intelligence-based methods. In this study, the number of normal events was significantly higher than the adventitious events. Also, among adventitious events, the dataset is unbalanced. Selecting the features based on this unbalanced dataset will significantly affect the performance of the final model. To overcome this challenge, we trained the neural networks used in this study utilizing the dynamic loss function. Moreover, we selected the final set of features based on a balanced dataset including the same number of normal and total adventitious events. In spite of all the techniques we employed to penalize the classes with fewer samples more intensely, the problem still needs to be investigated in much deeper detail in future works.

Another limitation of this research was that a considerable portion of events did not include the full breathing cycle the information of which, if present, could improve the performance of the proposed model. Also, in our auditory inspection, we noticed that some of the recorded signals included noise which also affects the performance of the model. Notwithstanding all the challenges, our proposed method could reach the total score of 0.74 which is superior to previous models [13, 19–23].

V. Conclusion

In this paper, multiple hybrid methods have been proposed to accomplish IEEE BioCAS Grand Challenge on Respiratory Sound Classification. The dataset has been prepared with different respiratory sound recordings from the lung in the pediatric population. To the best of our knowledge, this dataset is the biggest publicly available dataset on children which can support the applicability of the developed method to be used in this population. In contrast to the conventional methods which are either based on deep neural networks or manual feature extraction and analysis, our proposed framework integrates both methods to include not only the physiologically interpretable characteristics of the signals but also the data-driven features which are being missed in manual analyses. In the future, we will focus on preprocessing procedures to eliminate negatively impacting sources and try to tackle the imbalance data challenge using newer loss functions and other time-frequency representations such as wavelet transformation and S-transform.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. C. Jacome and A. Marques, "Computerized respiratory sounds in patients with copd: a systematic review," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, vol. 12, no. 1, pp. 104–112, 2015.
2. M. A. Islam, I. Bandyopadhyaya, P. Bhattacharyya, and G. Saha, "Multi-channel lung sound analysis for asthma detection," *Computer methods and programs in biomedicine*, vol. 159, pp. 111–123, 2018.
3. N. Montazeri Ghahjaverestan, S. Saha, M. Kabir, K. Zhu, B. Gavrilovic, and A. Yadollahi, "Estimating sleep apnea severity from tracheal signals using snore related features," in *TP131. TP131 NOVEL METHODOLOGIES FOR DIAGNOSING SLEEP DISORDERED BREATHING*. American Thoracic Society, 2021, pp. A4710–A4710.
4. A. Elwali and Z. Moussavi, "Obstructive sleep apnea screening and airway structure characterization during wakefulness using tracheal breathing sounds," *Annals of biomedical engineering*, vol. 45, no. 3, pp. 839–850, 2017.
5. Y. Ma, X. Xu, Q. Yu, Y. Zhang, Y. Li, J. Zhao, and G. Wang, "Lungbrn: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2019, pp. 1–4.
6. Y. Kim, Y. Hyon, S. S. Jung, S. Lee, G. Yoo, C. Chung, and T. Ha, "Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.
7. S. B. Shuvo, S. N. Ali, S. I. Swapnil, T. Hasan, and M. I. H. Bhuiyan, "A lightweight cnn model for detecting respiratory diseases from lung auscultation sounds using emd-cwt-based hybrid scalogram," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2595–2603, 2020.
8. J. S. Park, K. Kim, J. H. Kim, Y. J. Choi, K. Kim, and D. I. Suh, "A machine learning approach to the development and prospective evaluation of a pediatric lung sound classification model," *Scientific Reports*, vol. 13, no. 1, p. 1289, 2023.
9. A. Bohadana, G. Izbicki, and S. S. Kraman, "Fundamentals of lung auscultation," *New England Journal of Medicine*, vol. 370, no. 8, pp. 744–751, 2014.
10. M. Rosenthal and A. Bush, "The growing lung: normal development, and the long-term effects of pre-and postnatal insults," *European Respiratory Monograph*, vol. 7, pp. 1–24, 2002.
11. J. Schwartz, S. A. Katz, R. W. Fegley, and M. S. Tockman, "Sexand race differences in the development of lung function1-5," *Am Rev Respir Dis*, vol. 138, pp. 1415–1421, 1988.
12. Q. Zhang, J. Zhang, J. Yuan, H. Huang, Y. Zhang, B. Zhang, G. Lv, S. Lin, N. Wang, X. Liu *et al.*, "Grand challenge on respiratory sound classification for sprsound dataset," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2022, pp. 213–217.
13. ——, "Sprsound: Open-source sjtu paediatric respiratory sound database," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 5, pp. 867–881, 2022.
14. "Sprsound: Open-source sjtu paediatric respiratory sound database," 2022, data retrieved from GitHub, https://github.com/SJTU-YONGFU-RESEARCH-GRP/SPRSound.
15. P. Stasiakiewicz, A. P. Dobrowolski, T. Targowski, N. Gałazka-Świderek, T. Sadura-Sieklucka, K. Majka, A. Skoczylas, W. Lejkowski, and R. Olszewski, "Automatic classification of normal and sick patients with crackles using wavelet packet decomposition and support vector machine," *Biomedical Signal Processing and Control*, vol. 67, p. 102521, 2021.
16. D. Emmanouilidou, E. D. McCollum, D. E. Park, and M. Elhilali, "Computerized lung sound screening for pediatric auscultation in noisy field environments," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 7, pp. 1564–1574, 2017.
17. S. Li and Y. Liu, "Feature extraction of lung sounds based on bispectrum analysis," in *2010 Third International Symposium on Information Processing*. IEEE, 2010, pp. 393–397.
18. Ram Seshadri, "Featurewiz," 2020, data retrieved from GitHub, https://github.com/AutoViML/featurewiz.
19. J. Li, X. Wang, X. Wang, S. Qiao, and Y. Zhou, "Improving the resnet-based respiratory sound classification systems with focal loss," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2022, pp. 223–227.
20. L. Zhang, Y. Zhu, S. Tu, and L. Xu, "A feature polymerized based two-level ensemble model for respiratory sound classification," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2022, pp. 238–242.
21. W.-B. Ma, X.-Y. Deng, Y. Yang, and W.-C. Fang, "An effective lung sound classification system for respiratory disease diagnosis using densenet cnn model with sound pre-processing engine," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2022, pp. 218–222.
22. Z. Chen, H. Wang, C.-H. Yeh, and X. Liu, "Classify respiratory abnormality in lung sounds using stft and a fine-tuned resnet18 network," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2022, pp. 233–237.
23. N. Babu, J. Kumari, J. Mathew, U. Satija, and A. Mondal, "Multiclass categorisation of respiratory sound signals using neural network," in *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2022, pp. 228–232.