A Computationally Efficient Fully Convolutional Network for Respiratory Sound Classification

Ioanna Miliaresi
Department of Audio and Visual Arts
Ionian University
Corfu, Greece
imiliaresi@ionio.gr

Aggelos Pikrakis
Department of Informatics
University of Piraeus
Piraeus, Greece
pikrakis@unipi.gr

***Abstract*—This study introduces a novel, computationally efficient deep learning architecture for respiratory sound classification, designed for real-world medical applications. We propose a fully convolutional neural network that processes audio recordings of arbitrary length, ensuring flexibility in handling various input sizes. The network is evaluated on two classification tasks using the IEEE SPRSound Datasets of 2022 and 2023 Grand Challenges, leveraging mel spectrograms feature representations. Our architecture achieves high performance with a simple, low-cost design, making it adaptable to different classification formulations. The key innovation of our approach lies in balancing classification accuracy with minimal computational cost, enabling deployment in resource-constrained environments. The model demonstrates competitive performance compared to top-ranked methods, with compatibility for optimized computing environments, further enhancing efficiency and supporting its practical use in respiratory disease diagnosis and monitoring in clinical settings.**

***Index Terms*—Respiratory sound classification, deep learning, SPRSound dataset 2022 2023, BioCas challenge**

I. Introduction

Respiratory sound classification is essential for diagnosing and monitoring respiratory diseases by providing insights into lung function and identifying abnormalities. At the Tenth International Lung Sound Association (ILSA), respiratory sounds were classified into normal and abnormal categories. Abnormal sounds, such as crackles, wheezes, rhonchi, stridor, and pleural friction rubs, offer critical diagnostic information, with crackles, wheezes, and rhonchi being the most common.

Crackles, characterized by short, explosive, non-musical sounds, often signal parenchymal lung diseases such as pneumonia and pulmonary edema. Wheezes, described as musical high-pitched sounds, are commonly associated with airway diseases like asthma and Chronic Obstructive Pulmonary Disease. Rhonchi, resembling low-pitched snores, typically indicate the presence of airway secretions and are often cleared by coughing [1]. Adventitious sounds can be further classified into continuous adventitious sounds (CAS), which include rhonchi, wheezes, and stridor, and discontinuous adventitious sounds (DAS), such as coarse crackles and fine crackles, based on their duration [2].

Auscultation remains a valuable but variable diagnostic tool, with clinician experience influencing its effectiveness. Recent advancements in signal processing and machine learning have facilitated the automated classification of respiratory sounds, offering potential as a non-invasive assistive technology.

As reported in [3], significant efforts have been made to classify respiratory diseases, but they lack the design of a real-time hardware system that can automatically classify diseases based on symptoms, while maintaining low power consumption. Such a system is crucial for integration with multiple devices for automatic diagnosis.

In order to overcome this constraint, the main objective of our study is to formulate a simple deep learning architecture with low computational cost that achieves state-of-the-art classification results for respiratory sound classification. By designing an architecture that balances performance with computational efficiency, we seek to ensure that the model can be deployed in resource-constrained environments and utilized in real-time applications, ultimately enhancing the accessibility and reliability of respiratory health assessments. Several publicly available datasets have significantly contributed to advancements in respiratory sound analysis research. Notable datasets include the ICBHI 2017 Challenge Database, which focuses on detecting crackles and wheezes; the Pfizer21 2018 Database, containing samples of respiratory abnormalities such as coughing and sneezing; the Stethoscope22 2021 Database, which consists of lung sound recordings; and the HF Lung V123 2021 Database, which provides respiratory recordings along with demographic information. More recently, the IEEE Grand Challenges on Respiratory Sound Classification introduced the SPRSound Dataset [4]. Our study focuses on the SPRSound 2022 and 2023 datasets, as they enable a direct performance comparison with the top-performing challenge methods using standardized evaluation metrics.

The IEEE challenge includes two tasks:

- **Task 2-1 Ternary**: A 3-class task for classifying respiratory sound records into the categories of “Normal”, “Adventitious”, and “Poor Quality”.
- **Task 2-2 Multi-class**: A 5-class task for classifying respiratory sound records into the categories of “Normal”, “CAS”, “DAS”, “CAD & DAS”, and “Poor Quality”.

The top-rated method of the 2022 challenge [5] used a fixed-length segmentation scheme, where spectrogram segments were input into a ResNet-based classifier. The study introduced two ResNet architectures: the original ResNet and a temporal convolutional TC-ResNet. During preprocessing, segmentation and zero-padding were applied for audio samples shorter than the specified length, with random time shift padding during training and centered padding during testing. Any excess audio beyond the specified length was removed during testing.

The top-rated method of the 2023 challenge, [6] applied several preprocessing techniques, including pre-normalization for scaling respiratory sounds on Mel frequency cepstrum coefficients (MFCCs). They applied random flipping and cropping. The processed spectrograms were fed into a pre-trained supervised contrastive model and encoded into high-dimensional embeddings.

In contrast, our novel approach eliminates the necessity for segmentation schemes by considering each respiratory sound recording in its entirety, treating it as a 2-D representation (image) directly analyzed by a fully convolutional neural network architecture. This enables our method to handle sounds of arbitrary duration without requiring segmentation as a preprocessing stage, thereby preserving the temporal characteristics of the sounds and simplifying the analysis pipeline. Our system not only offers a more straightforward and comprehensive approach but also maintains performance comparable to existing methods, while ensuring a low computational cost.

The proposed architecture boasts a low computational cost due to its minimal number of model parameters and weights. Consequently, the architecture can be effectively executed on standard GPU hardware, where execution time is remarkably low. This efficiency is evidenced by our timing analysis, which demonstrates that the model runs swiftly on a typical GPU, making it an attractive option for resource-constrained environments and real-time applications.

Through the following sections, we first review existing research on respiratory sound classification in Section II. We then present our method in Section III, describe the available datasets, and explain our feature extraction, model design, and training processes. Section IV presents the experimental results, followed by the conclusions in Section V.

II. Related work

Recent literature indicates that frequency domain analysis is particularly well-suited for the classification of respiratory sounds. The spectral characterization of these sounds primarily relies on MFCC coefficients [[7], [8]] and variations of spectrograms [[9], [10], [11], [12]], which have been shown to yield high classification accuracy.

The most notable advancements are associated with deep learning techniques. More specifically, Razvadauskas et al. [8] explored supervised models leveraging tree-based ensemble methods. Cozzatti et al. [13] proposed a weakly supervised approach based on a Variational Autoencoder. Shuvo et al. [14] introduced a lightweight CNN architecture that operates on features derived from empirical mode decomposition and the continuous wavelet transform. Pham et al. [15] focused on scalogram representations and CNNs. Ntalampiras et al. [16] proposed a Siamese Neural Network framework while, Bae et al. [11] introduced patch-mix contrastive learning with an audio spectrogram transformer. Pessoa et al. [12] proposed a dual-input deep learning architecture, leveraging raw audio signals and STFT spectrograms. Lal et al. [17] employed transfer learning with a VGGish-stacked BiGRU model. Meanwhile, Yang et al. [18] introduced BLNet, which integrated ResNet, GoogleNet, and self-attention mechanisms. Kim et al. [1] explored the VGG architecture.

Since we experimented with the 2022 and 2023 IEEE BioCas Grand Challenges data we focus on their top-rated works. To be more detailed, Li et al. [5] proposed an ResNet-based respiratory sound classification system with ResNet18 and TC-ResNet algorithms and achieved top scores, including the best ternary and multi-class scores of 0.833 and 0.673. Zhang et al. [19] introduced a feature-polymerized two-level ensemble model while Ma et al. [20] proposed a DenseNet169 CNN model with optimized preprocessing methods, achieving classification scores of 0.838 and 0.673 for the two tasks. Chen et al. [21] compared the performance of different feature extraction techniques, including STFT, Mel spectrograms, and Wav2vec, and employed pre-trained ResNet18, LightCNN, and audio spectrogram transformer algorithms, achieving notable harmonic scores of 0.71 and 0.36 for the two tasks. Babu et al. [22] proposed a convolution-based deep learning model using MFCCs, achieving a score of 0.876 and 0.515 for ternary and multiclass tasks.

As far as the top three announced works for the 2023 challenges are concerned [6], [23], and [12] all contribute with deep learning approaches. Hu et al. focus on addressing class imbalance by applying supervised contrastive pretraining. They used MFCCs, pre-normalization, and spectrograms, and applied a supervised contrastive model Their approach achieved a ternary score of 0.8097 and a multiclass score of 0.6666. Ngo et al. excelled with spatio-temporal modeling, using continuous wavelet transformation for feature extraction and various data augmentation techniques, such as spectrogram oversampling, masking, and mixup. They employed an inception-residual network with spatio-temporal focusing and multi-head mechanisms, achieving a ternary score of 0.7693 and a multiclass score of 0.6318. Finally, Pessoa et al. proposed a dual-input deep learning architecture, utilizing both raw audio signals and STFT spectrograms, processed through separate CNN blocks. This approach achieved a ternary score of 0.756 and a multiclass score of 0.4666.

III. METHOD DESCRIPTION

A. DATASET DESCRIPTION

For our experiments, we utilised the SPRSound Open-Source SJTU Pediatric Respiratory Sound Database [24], which includes respiratory sounds from children aged 1 month to 18 years, recorded at the Shanghai Children's Medical Center using a Yunting Model II stethoscope. The 2022 dataset was partitioned into training and test sets. The training set consists of $1,949$ records containing $6,656$ respiratory sound events collected from 251 participants, while the test dataset

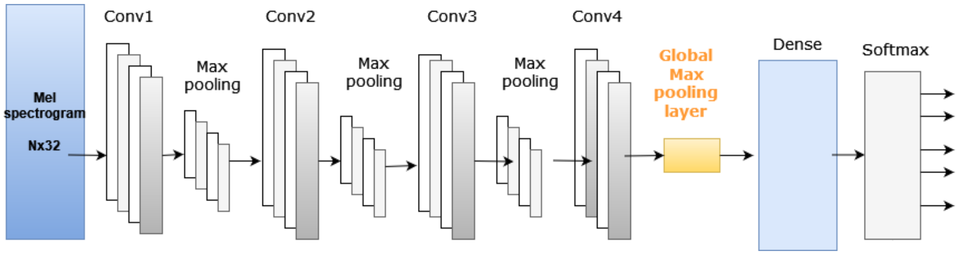

Fig. 1. The network architecture, implemented as a fully convolutional neural network

comprises 356 records. The 2023 partition included a training set of 1,949 records and a testing set of 870 records.

B. FEATURE EXTRACTION PROCESS

At the first stage of the processing pipeline, a short-term feature extraction method is employed to derive a sequence of vectors of MFCCs from the input signal. The signal is first amplitude-normalized and resampled to 44.1 kHz. At each frame, the Discrete Fourier Transform (DFT) is computed, and the DFT coefficients are fed into a mel filter bank. Each mel filter implements a weighted sum of the magnitudes of the DFT coefficients within its frequency range. The logarithm of each filter bank output is then computed, followed by the DCT. Only the first 13 MFCCs are retained. This results in a 2-D representation (image) of the input segment, with dimensions $N \times 13$, where N varies depending on the duration of the audio recording.

Additionally, we extract the mel spectrogram of each audio signal. For the sake of consistency, all signals are again amplitude-normalized and resampled at 44.1 kHz. Subsequently, the spectrogram is computed using a 1024-point DFT and a periodic Hamming window with a duration of 40 ms and a 25 ms overlap. The spectrogram is then passed through a mel-scale filter bank of 32 filters, covering the frequency range from 20 Hz to 20 kHz, i.e., the entire audible spectrum. The resulting feature sequence is again represented as a 2-D image with dimensions $N \times 32$, where N varies depending on the duration of the audio recording.

C. NEURAL NETWORK ARCHITECTURE

The novelty of our approach lies in leveraging a network architecture capable of handling audio recordings of arbitrary duration. Unlike conventional methods, we refrain from utilizing segmentation or zero-padding procedures. This choice is dictated not only by the key features of our architecture but also by our observations that segmentation schemes tend to overlook important information regarding the temporal evolution of audio features across the entire sample. This is made possible by a fully convolutional neural network (see [25] for a definition of fully convolutional architectures). The key property of such neural network architectures is their ability to adeptly process inputs of varying dimensions while yielding an output vector of fixed dimensionality

In a standard convolutional classifier, a convolutional layer is typically followed by a pooling layer, and this design pattern is repeated several times. The feature maps from the last convolutional block are then flattened and passed through a cascade of fully connected layers to generate the final prediction outcome, e.g., a classification decision. It therefore follows that the input images must have a fixed size to ensure a consistent number of inputs at the first dense layer of the feed-forward part.

To overcome this limitation, we use fully convolutional networks, a technique newly introduced for arbitrary-size splitting augmentation in our previous work [26] and detailed in [27].

In this method, the final convolutional block is replaced with a block consisting of a convolutional layer with a kernel size of 1×1 and a stride of 1, followed by a global max pooling layer. The output dimensionality of this block remains consistent, determined solely by the number of filters, denoted as n, resulting in an output size of $1 \times 1 \times n$. Notably, this dimensionality remains invariant to changes in the input image size and is subsequently forwarded to the fully connected layer.

In our method, we represent each audio recording as a single-channel, two-dimensional ”image” with dimensions defined by its height h and width w. When extracting MFCCs features, this results in a matrix with N rows and 13 columns. Similarly, for mel spectrogram representations, the matrix size becomes N rows by 32 columns. In this methodology, N signifies the number of frames extracted from the audio signal, which depends on the recording's duration. This consistency applies to both MFCCs and mel spectrogram representations. Typically, N ranges between 124 and 1462 frames, accommodating various durations commonly found in respiratory sound recordings. As a consequence, the input shape of the first convolutional layer does not have fixed dimensions, and having adopted a batch size equal to one, the resulting batch shape is $1 \times N \times 32 \times 1$. More specifically, as illustrated in Figure 1, the proposed architecture consists of:

- Four consecutive convolutional layers. Each layer contains 64, 64, 32, and 32 convolutional masks, respectively. The first three have a kernel size of 3×3. The final convolutional layer has a kernel size of 1 and is followed by a global max pooling layer. The output of each convolutional operation is processed through a Rectified Linear Unit (ReLU) function, and the resulting feature matrix is subsequently subsampled by a max pooling layer with a size of 2×2. Each 52×32 input matrix

TABLE I
TEST RESULTS FOR DIFFERENT FEATURE REPRESENTATIONS ON THE FIVE-CLASS AND THREE-CLASS TASKS.

Experimental results referring to 2022 SPRSound data					
Task 2-1: Ternary-class classification					
Input features	Accuracy (%)	Specificity	Sensitivity	Average_score	Harmonic_score
MFCCs	0.76 ± 0.07	0.75 ± 0.05	0.91 ± 0.05	0.75 ± 0.06	0.75 ± 0.04
Mel Spectrogram	0.86 ± 0.04	0.84 ± 0.03	0.86 ± 0.03	0.86 ± 0.04	0.86 ± 0.02
Task 2-2: Multi-class classification					
Input features	Accuracy (%)	Specificity	Sensitivity	Average_score	Harmonic_score
MFCCs	0.575 ± 0.03	0.5 ± 0.05	0.90 ± 0.02	0.57 ± 0.03	0.57 ± 0.03
Mel Spectrogram	0.625 ± 0.05	0.62 ± 0.06	0.95 ± 0.04	0.62 ± 0.06	0.62 ± 0.05

Experimental results referring to 2023 SPRSound data					
Task 2-1: Ternary-class classification					
Input features	Accuracy (%)	Specificity	Sensitivity	Average_score	Harmonic_score
MFCCs	0.61 ± 0.03	0.55 ± 0.04	0.83 ± 0.02	0.75 ± 0.03	0.75 ± 0.05
Mel Spectrogram	0.72 ± 0.02	0.69 ± 0.03	0.58 ± 0.04	0.68 ± 0.03	0.63 ± 0.03
Task 2-2: Multi-class classification					
Input features	Accuracy (%)	Specificity	Sensitivity	Average_score	Harmonic_score
MFCCs	0.66 ± 0.02	0.61 ± 0.03	0.72 ± 0.02	0.67 ± 0.02	0.65 ± 0.02
Mel Spectrogram	0.68 ± 0.01	0.72 ± 0.02	0.65 ± 0.03	0.68 ± 0.02	0.67 ± 0.02

produced by the preprocessing stage is passed through the convolutional layers.

- Subsequently, the outputs of the global max pooling layer are forwarded through a dense layer comprising 128 neurons with a ReLU activation function.
- Finally, a softmax layer with five outputs calculates the final classification decision.

The above description refers to a five-class task. The configuration was modified to accommodate the three-class problem. This modification only required changing the output layer to use a softmax function with three outputs instead of five.

IV. Experiments and results

To assess the performance of the classifier, we conducted experiments based on the metrics introduced by the challenges to facilitate comparison with the top-ranked methods. Therefore, we evaluated the performance in terms of sensitivity (SE), specificity (SP), average score (AS), and harmonic score (HS). We aimed to assess the performance of two audio features to determine the most effective one. Building on previous research, we evaluated our neural network architecture using MFCCs and the mel spectrogram across both classification tasks.

The proposed classifier was trained for 500 epochs using the Adam gradient descent algorithm and a fixed learning rate of 0.0001 to optimize the standard cross-entropy loss function. The training process followed a 3-fold cross-validation scheme. As is standard practice, an early stopping criterion was adopted and a dropout regularization scheme of 0.5 was applied to the convolutional layers.

We experimented with the two tasks under study: the ternary classification task (Task 2-1) and the five-class classification task (Task 2-2). The results are summarized in Table **??** and indicate competitive performance for both feature representations on both datasets.

More specifically, for ternary classification on the 2022 dataset, the model demonstrates competitive performance for both feature representations. The MFCCs-based approach achieved an accuracy of 0.76 ± 0.07, with a sensitivity of 0.91 ± 0.05 and a specificity of 0.75 ± 0.05. Similarly, the mel spectrogram representation yielded a higher accuracy of 0.86 ± 0.04, with comparable sensitivity and specificity values of 0.86 ± 0.03 and 0.84 ± 0.03, highlighting its effectiveness for the classification task.

In the case of multi-class classification, performance was evaluated in a similar manner. As shown in Table **??**, the MFCCs-based approach achieved an accuracy of 0.575 ± 0.03, with a sensitivity of 0.90 ± 0.02 and a specificity of 0.50 ± 0.05. The mel spectrogram representation outperformed MFCCs, achieving an accuracy of 0.625 ± 0.05 and demonstrating improved sensitivity (0.95 ± 0.04) and specificity (0.62 ± 0.06) scores.

Focusing on the 2023 data, for ternary classification, the MFCCs-based scheme achieved an accuracy of 0.61 ± 0.03, with a sensitivity of 0.83 ± 0.02 and a specificity of 0.55 ± 0.04. The mel spectrogram approach performed better, reaching an accuracy of 0.72 ± 0.02, with balanced sensitivity (0.58 ± 0.04) and specificity (0.69 ± 0.03) values.

Regarding multi-class classification, MFCCs obtained an accuracy of 0.66 ± 0.02, with a sensitivity of 0.72 ± 0.02 and a specificity of 0.61 ± 0.03. Meanwhile, the mel spectrogram representation further improved classification performance, achieving an accuracy of 0.68 ± 0.01, with sensitivity (0.65 ± 0.03) and specificity (0.72 ± 0.02) values.

Our experimentation with GPU acceleration, specifically leveraging TikTok, demonstrated notable enhancements in the computational efficiency of the task. The device utilized had the following properties: pciBusID:0000 : 21 : 00.0, name: NVIDIA GeForce RTX 2080 Ti, compute capability: 7.5, core clock: $1.545GHz$, core count: 68, device memory size: $10.76GiB$, and device memory bandwidth: 573.69 GiB/s. The evaluation time for the 2022 and 2023 testing datasets was measured at $31,569.88$ ms and $59,012.28$ ms, respectively. Notably, the model comprises a small number of parameters: $29,006$ in total, with $28,622$ trainable and 384 non-trainable parameters.

These results highlight that a fully convolutional network processing audio recordings at arbitrary sizes without preprocessing can effectively operate in the task of respiratory sound classification. While both MFCCs and the mel spectrogram exhibit satisfactory performance, the mel spectrogram outperforms.

V. CONCLUSIONS

Our experimental study indicates that processing recordings at their original duration, without segmentation, presents a viable approach to addressing the complex challenge of respiratory sound classification. Specifically, we demonstrate the effectiveness of employing mel spectrogram representations of audio signals within a fully convolutional neural network architecture, establishing a competitive method within the IEEE BioCAS challenge framework. These findings underscore the potential of leveraging full-length recordings and optimized feature representations to enhance respiratory sound classification, contributing to the development of more accurate diagnostic and monitoring tools in respiratory health. Our future plans include refining our proposed architecture and expanding our experiments to all available respiratory databases.

ACKNOWLEDGEMENTS

This work has been partly supported by the University of Piraeus Research Center.

REFERENCES

[1] Y. Kim *et al.*, "Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning," *Scientific reports*, vol. 11, no. 1, p. 17186, 2021.

[2] T. Xia *et al.*, "Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues," *Experimental Biology and Medicine*, vol. 247, no. 22, pp. 2053–2061, 2022.

[3] A. F. Mahmood, A. M. Alkababji, and A. Daood, "Resilient embedded system for classification respiratory diseases in a real time," *Biomedical Signal Processing and Control*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266711173

[4] Z. Qing *et al.*, "Grand challenge on respiratory sound classification for sprsound dataset," *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–5, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:253556967

[5] L. Jun *et al.*, "Improving the resnet-based respiratory sound classification systems with focal loss," *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 223–227, 2022.

[6] N. Dat *et al.*, "A deep learning architecture with spatio-temporal focusing for detecting respiratory anomalies," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2023, pp. 1–5.

[7] C. Papadakis *et al.*, "Ausculnet: A deep learning framework for adventitious lung sounds classification," *2023 30th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 1–4, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266905400

[8] H. Razvadauskas *et al.*, "Exploring traditional machine learning for identification of pathological auscultations," *arXiv preprint arXiv:2209.00672*, 2022.

[9] T. K. T. Nguyen and F. Pernkopf, "Lung sound classification using co-tuning and stochastic normalization," *IEEE Transactions on Biomedical Engineering*, vol. PP, pp. 1–1, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236912776

[10] T. T. Nguyen and F. Pernkopf, "Lung sound classification using snapshot ensemble of convolutional neural networks," *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 760–763, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221387767

[11] B. Sangmin *et al.*, "Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification," *ArXiv*, vol. abs/2305.14032, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258841333

[12] D. Pessoa *et al.*, "Pediatric respiratory sound classification using a dual input deep learning architecture," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2023, pp. 1–5.

[13] M. Cozzatti, F. Simonetta, and S. Ntalampiras, "Variational autoencoders for anomaly detection in respiratory sounds," in *International Conference on Artificial Neural Networks*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251402903

[14] H. B. Samiul, Based Shuvo, "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, pp. 2595–2603, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221556950

[15] T. V. Pham *et al.*, "Classification of lung sounds using scalogram representation of sound segments and convolutional neural network," *Journal of Medical Engineering & Technology*, vol. 46, no. 4, pp. 270–279, 2022.

[16] S. Ntalampiras, "Explainable siamese neural network for classifying pediatric respiratory sounds," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[17] K. N. Lal, "A lung sound recognition model to diagnoses the respiratory diseases by using transfer learning," *Multimedia Tools and Applications*, pp. 1–17, 2023.

[18] R. Yang *et al.*, "Respiratory sound classification by applying deep neural network with a blocking variable," *Applied Sciences*, vol. 13, no. 12, p. 6956, 2023.

[19] Z. Lin *et al.*, "A feature polymerized based two-level ensemble model for respiratory sound classification," *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 238–242, 2022.

[20] M. Weijie *et al.*, "An effective lung sound classification system for respiratory disease diagnosis using densenet cnn model with sound pre-processing engine," *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 218–222, 2022.

[21] C. Zizhao *et al.*, "Classify respiratory abnormality in lung sounds using stft and a fine-tuned resnet18 network," *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 233–237, 2022.

[22] B. Naseem *et al.*, "Multiclass categorisation of respiratory sound signals using neural network," *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 228–232, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253555602

[23] J. Hu *et al.*, "Supervised contrastive pretrained resnet with mixup to enhance respiratory sound classification on imbalanced and limited dataset," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2023, pp. 1–5.

[24] Q. Zhang *et al.*, "Sprsound: Open-source sjtu paediatric respiratory sound database," *IEEE Trans. on Biomedical Circuits and Systems*, vol. 16, no. 5, pp. 867–881, 2022.

[25] J. Long *et al.*, "Fully convolutional networks for semantic segmentation," in *Procs. of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[26] I. Miliaresi and A. Pikrakis, "A modular deep learning architecture for voice pathology classification," *IEEE Access*, vol. 11, pp. 80 465–80 478, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260411537

[27] I. Miliaresi, "Digital audio processing methods for voice pathology detection," Ph.D. dissertation, University of Piraeus, 2025. [Online]. Available: https://theses.eurasip.org/theses/1010/digital-audio-processing-methods-for-voice/