A Non-Linear Triple-Stream Self-Operational Framework for Respiratory Sound Classification

Hoang Vo[1], *Thien Quoc* Nguyen[1], *Hoang Long* Nguyen Van[1], *Quang Thang* Ngo[1], *Quang Tan* Nguyen[1], *Ho-Si-Hung* Nguyen[1,*], and *Hai Canh* Vu[2]

[1]University of Science and Technology - The University of Danang, Da Nang, Vietnam
[2]Roberval Laboratory, University of Technology of Compiegne, France

Abstract. Automated respiratory sound classification is a critical component for the early diagnosis of chronic pulmonary diseases; however, traditional deep learning architectures, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, are often limited by the linear nature of their convolutional kernels. These models frequently struggle to capture the highly non-linear, explosive transients characteristic of adventitious sounds like crackles and wheezes. This paper proposes a novel Triple-Stream Self-Operational Neural Network (SelfONN) framework designed to overcome these linear constraints. The architecture introduces a Parallel Spectral Stream Extraction (PSSE) module that extracts and fuses heterogeneous time-frequency representations—specifically Mel Spectrograms, Mel-Frequency Cepstral Coefficients (MFCC), and Constant-Q Transform (CQT)—preserving spatial dimensions to prevent temporal blurring. By replacing standard linear convolutions with high-order cubic operational kernels ($Q = 3$ and $Q = 4$), the network natively approximates complex acoustic signatures with a shallow yet highly non-linear architecture. Evaluated on the ICBHI 2017 Scientific Challenge dataset using 10-fold stratified cross-validation and Focal Loss to mitigate class imbalance, the proposed model achieves a state-of-the-art accuracy of **92.23%** and an F1-score of **92.26%**. These results demonstrate a significant improvement over existing CNN and hybrid RNN baselines.

1 Introduction

Respiratory diseases are among the leading causes of mortality worldwide, contributing significantly to global healthcare burden [1]. A key clinical indicator of these diseases is the presence of abnormal lung sounds, commonly referred to as adventitious sounds, such as *wheezes* and *crackles* [2]. Wheezes are continuous, musical sounds typically associated with airway obstruction, while crackles are discontinuous, explosive sounds often linked to alveolar abnormalities [3]. These acoustic signatures provide essential diagnostic information for diseases such as pneumonia, asthma, and chronic obstructive pulmonary disease (COPD) [4].

In clinical practice, physicians rely on auscultation using a stethoscope to detect these abnormal sounds. However, this approach suffers from several limitations. Gurung *et al.* [5] demonstrated that traditional auscultation is highly dependent on clinician expertise and lacks reproducibility. Furthermore, McCollum *et al.* [6] highlighted that environmental noise and recording conditions can significantly degrade auscultation reliability, making it difficult to detect subtle abnormal sounds such as wheezes and crackles.

These limitations motivate the need for automated systems capable of objectively analyzing respiratory sounds.

*e-mail: nhshung@dut.udn.vn

Recent benchmarking on the ICBHI 2017 dataset highlights the performance bottlenecks of standard AI architectures. Early machine learning approaches, such as the Support Vector Machine (SVM) framework proposed by Chambres et al. [7], established initial baselines for adventitious sound classification but exhibited low accuracy (approximately 49.98%) due to their reliance on manually engineered features and limited capacity for modeling complex acoustic patterns.

Transitioning to deep learning, researchers leveraged convolutional architectures to process time-frequency representations. For instance, Gairola et al. [8] introduced RespireNet, a ResNet-34 based model, to detect abnormal lung sounds. While residual connections mitigated vanishing gradients, the performance of standard CNNs remains constrained by the linear $W * x$ convolutional operation. To capture temporal variations within noisy respiratory cycles, hybrid CNN-RNN models have been widely adopted. Petmezas et al. [9] implemented a hybrid CNN-LSTM network coupled with a focal loss function, achieving a 10-fold cross-validation accuracy of 76.39% for 4-class categorization.

More recently, researchers have explored waveform-based encoders to bypass the limitations of fixed time-frequency resolutions. Kulkarni et al. [10] utilized a SincNet-based architecture combined with self-supervised contrastive pretraining, achieving a high accuracy of

90.60%. While SincNet improves the extraction of low-level features directly from raw audio, such models still struggle with the severe class imbalance inherent in the ICBHI dataset, frequently failing to distinguish the rare "Both" class (concurrent crackles and wheezes).

Furthermore, the vast majority of these models rely on singular feature inputs, predominantly Mel-spectrograms, STFT, or raw waveforms. This homogeneous feature representation often obscures the sharp, low-frequency transients essential for crackle detection, a critical oversight in existing single-stream pipelines.

Moreover, standard CNN [11] and LSTM [12] architectures primarily rely on linear transformations followed by scalar activation functions. This linear kernel formulation struggles to accurately approximate the highly non-linear, burst-like transients of fine crackles without deploying excessively deep networks, which in turn leads to overfitting on limited medical datasets. To bridge this gap, we introduce a non-linear operational paradigm.

To address the feature homogeneity of above frameworks and the linear constraints of CNNs and LSTM, we propose a novel framework with the following contributions:

1. **Parallel Spectral Stream Extraction (PSSE):** We transition away from single-feature inputs by extracting Mel, MFCC, and CQT representations in parallel, preserving the natural spatial dimensions to prevent temporal blurring.

2. **High-Order Self-Operational Kernels:** We replace standard linear CNN convolutions with Self-Operational Neural Networks (SelfONN) utilizing a cubic operational order ($Q = 3$ and $Q = 4$), allowing the network to natively approximate non-linear acoustic signatures.

3. **High Performance:** By combining PSSE, Self-ONN, and Focal Loss, we establish a new benchmark of 92% accuracy on the ICBHI 2017 dataset, robustly outperforming standard CNN, SVM, and LSTM baselines.

The paper is organized as follows: Section II covers the proposed methodology and training database. Sections III provides a comprehensive experiment results of the framework and comparison with other models. Section IV is the conclusion of our paper, which also proposes future exploration potentials on this topic.

2 Proposed Methodology

To address the complexities of respiratory acoustic signals, we propose a **Triple-Stream Self-Operational Neural Network**. This architecture is specifically designed to extract, process, and fuse heterogeneous time-frequency representations without compromising natural spatial dimensions.

2.1 Training database

The proposed system is developed and evaluated using the ICBHI 2017 Scientific Challenge dataset [13]. This dataset comprises 920 annotated lung sound recordings collected from 120 patients across diverse age groups, ranging from infants to the elderly. To simulate real-world clinical environments, the database includes both clear-cut respiratory sounds and noisy recordings captured across multiple clinical centers using various stethoscopes (e.g., Meditron, Littmann 3200).

In total, the dataset contains approximately 5.5 hours of audio, encompassing 6,898 annotated respiratory cycles. These cycles are categorized into four pathological states based on the presence of adventitious sounds: 1,864 cycles contain crackles, 886 contain wheezes, 506 contain both crackles and wheezes, and the remainder represent normal healthy cycles.

2.2 Preprocessing and Augmentation

Audio signals are inherently susceptible to ambient noise and physiological artifacts. All recordings are first resampled to 16 kHz to standardize the input while preserving the Nyquist frequency limit sufficient for respiratory acoustics. A 12th-order Butterworth bandpass filter is applied to eliminate low-frequency cardiac interference and high-frequency ambient noise. The signal is subsequently normalized to a $[-1, 1]$ amplitude range.

To maximize training diversity and mitigate the severe class imbalance, we extract overlapping 3.6-second time windows with a 1.8-second stride (50% overlap). Furthermore, a two-tier data augmentation strategy is employed:

- **Offline VTLP:** Vocal Tract Length Perturbation (VTLP) [14] is applied to simulate variations in patient body habitus. VTLP used a warping factor α in the range 0.82–1.18, transition frequency $f_{\text{hi}} = 2000$ Hz, loudness scaling $\lambda = 0.9$–1.1, and Gaussian noise with standard deviation 0.005.
- **Online SpecAugment:** Aggressive dynamic masking [15] is applied during training to enforce robust feature learning and prevent overfitting. Online SpecAugment used a time-warping distance of 3, two frequency masks with a maximum width of 17 bins, and two time masks with a maximum width of 10 time steps.

2.3 Parallel Spectral Stream Extraction

Respiratory pathologies exhibit distinct acoustic signatures. Rather than relying on a single representation, our framework use **Parallel Spectral Stream Extraction (PSSE)** which contains three independent branches of frequency domain feature. To prevent temporal resolution loss, these features preserve their natural spatial dimensions.

Mel Spectrogram:

The Mel spectrogram maps the linear frequency spectrum to the non-linear Mel scale [16], mimicking human pitch

perception to effectively isolate rhythmic breathing from chaotic pathological noise [17]. We compute 64 Mel-frequency bins using a 2048-sample FFT window and 1024-sample hop length to capture the broad spectral morphology of the lung signal.

Mel-Frequency Cepstral Coefficients (MFCC):

MFCCs represent the spectral envelope by applying a Discrete Cosine Transform (DCT) to log-Mel filterbank outputs. This de-correlates the coefficients to isolate the "filter" characteristics of the respiratory tract [18], providing a robust signature of structural airway dimensions [19]. We extract the first 40 coefficients to ensure a high signal-to-noise ratio for the classification layers.

Constant-Q Transform (CQT):

Unlike the fixed-resolution STFT, the CQT [20] maintains a constant Q-factor ($Q = f_k/\Delta f_k$), providing multi-resolution analysis. This allows for high frequency resolution of low-frequency harmonics (wheezes) and high temporal resolution for impulsive transients (crackles) [21]. Our implementation utilizes 64 bins (12 per octave) starting at 50 Hz to prevent the temporal blurring of fine-grained transients.

2.4 Model Framework

The core of the model framework leverages 2D Self-Operational Neural Networks (SelfONN) [22]. Unlike traditional convolutional layers that utilize a single linear kernel, a SelfONN layer with an operational order Q performs a sum of convolutions over successive powers of the input:

$$y = \sum_{n=1}^{Q} (W_n * x^n) + b \tag{1}$$

where $*$ denotes the convolution operator, W_n represents the learnable filter weights for the n-th order, and x^n is the input feature map raised to the n-th power via element-wise multiplication.

As illustrated in Fig. 1, the model features three independent streams corresponding to the Mel, MFCC, and CQT inputs. Each stream consists of three sequential SelfONN blocks with increasing channel depths ($24 \rightarrow 48 \rightarrow 128$). Each block incorporates SelfONN, stream dropout (0.05), Batch Normalization, ReLU activation, and max pooling.

Because the PSSE inputs possess variable spatial dimensions, the GAP layer is used to generate the 1-D output embedding from each of the stream. The resulting 128-dimensional vectors from each stream are concatenated into a single 384-dimensional fused feature vector which is then passed through the ReLu activation layer and spatial dropout layer with a dropout factor of 0.1. This joint representation is passed through two dense layers (160 and 80 neurons) with spatial dropout before the final 4-way softmax classification. To directly address the severe class imbalance (particularly the rare “Both” class), the network is optimized using Focal Loss [23], ensuring the model heavily penalizes misclassifications on hard examples.

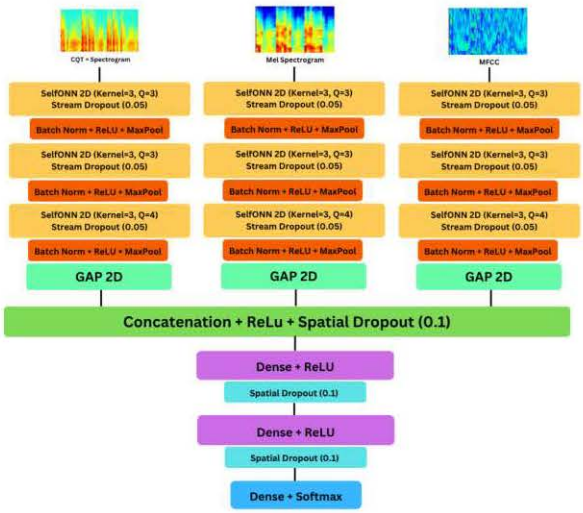

Figure 1. Our Model Architecture. Three parallel SelfONN streams process the PSSE representations. Features are fused via concatenation and passed through dense classification layers.

2.5 Focal Loss for Pathological Imbalance

A significant challenge in respiratory sound classification using the ICBHI 2017 dataset is the inherent class imbalance. In Fig. 2, the dataset is dominated by normal respiratory cycles, while pathological instances—especially the "Both" class (concurrent crackles and wheezes)—are underrepresented. Standard Categorical Cross-Entropy (CCE) loss treats all misclassifications equally, which often leads the model to become biased toward the majority class, resulting in poor sensitivity for rare clinical events.

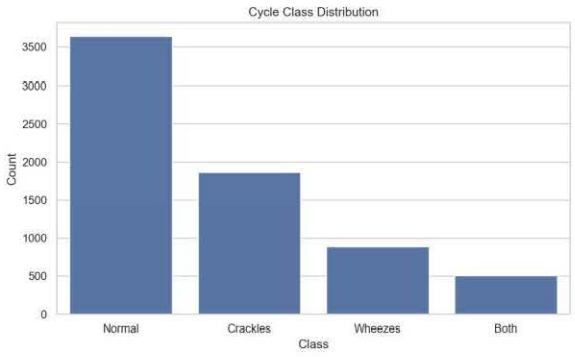

Figure 2. ICBHI 2017 classes profile

To mitigate this, we employ Focal Loss [23], which reshapes the standard cross-entropy loss by adding a modulating factor that down-weights the loss contribution from "easy" (well-classified) examples and focuses training on "hard" pathological transients. The Focal Loss (FL) is mathematically defined as:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{2}$$

where p_t is the model's estimated probability for the target class. The loss function incorporates two key hyperparameters:

- **Focusing Parameter (γ):** We set $\gamma = 2.0$. As γ increases, the modulating factor $(1 - p_t)^\gamma$ reduces the relative loss for well-classified examples ($p_t > 0.5$), effectively forcing the SelfONN kernels to prioritize the subtle, non-linear features of misclassified crackles and wheezes.
- **Weighting Factor (α):** An α-balanced variant is used to handle the specific prior probabilities of the four classes, ensuring that the gradient updates are not overwhelmed by the high frequency of normal cycles.

By integrating Focal Loss with the cubic operational kernels of the SelfONN, the framework achieves a high degree of diagnostic robustness, ensuring that the 92% Accuracy and F1-score are balanced across all pathological categories rather than being inflated by the majority class.

3 Experimental Results

3.1 Setup and Dataset

All experiments were implemented in PyTorch and trained on an NVIDIA GeForce RTX 3080 Ti GPU with CUDA 12.6. The model was evaluated using 10-fold stratified cross-validation. Patients in nine folds were used for training, while patients in the remaining fold were used for validation. Data augmentation was applied only to the training set, and the validation set was evaluated without augmentation.

The model was trained with a batch size of 40 for a maximum of 250 epochs. Adamax was used as the optimizer with an initial learning rate of 0.0003, weight decay of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. A ReduceLROnPlateau scheduler was applied with a reduction factor of 0.7 and a patience of 10 epochs. Early stopping was used with a patience of 35 epochs, and the best model checkpoint in each fold was selected according to validation accuracy.

3.2 10-Fold Stratified Cross-Validation

Our study adopted a patient-independent 10-fold stratified cross-validation protocol. First, each subject in the ICBHI 2017 dataset was identified using the patient ID provided in the recording metadata. The fold assignment was then performed at the patient level rather than at the respiratory-cycle or window level. As a result, all recordings belonging to the same patient were assigned to a single fold only.

After the patient-level folds were created, the audio preprocessing and segmentation steps were performed separately within each fold. For each validation round, patients from nine folds were used for training, while patients from the remaining fold were held out for validation. The 3.6-second windows with a 1.8-second stride were generated only after this split, ensuring that overlapping segments from the same recording could not appear in both the training and validation sets.

All augmentation methods, including VTLP and SpecAugment, were applied only to the training folds. As a result, the validation data always came from unseen patients and was not used during training or augmentation.

3.3 Classification Performance

The overall robustness of the model framework is summarized in Table 1. The model achieved an aggregate Accuracy of 92.23% and a corresponding F1-Score of 92.26%. Notably, the standard deviation across the 10-fold cross-validation remained exceptionally low (±0.0046), indicating high model stability and consistent performance across different patient subsets. The precision (92.33%) and recall (92.23%) values are closely aligned, suggesting a well-balanced model that does not exhibit a bias toward either type I or type II errors.

Table 1. Aggregate Performance Metrics (Mean ± SD) over 10-Fold Cross-Validation.

Metric	Value (%)
Accuracy	92.23 ± 0.0046
F1-Score	92.26 ± 0.0046
Precision	92.33
Recall	92.23

To evaluate the model's ability to distinguish between specific respiratory pathologies, a detailed classification report is provided in Table 2. The results demonstrate high discriminatory power across all four categories:

- **Adventitious Sound Detection:** The model showed superior performance in detecting *Wheezes* and the complex *Both* (Crackle and Wheeze) class, achieving F1-scores of 0.9644 and 0.9510, respectively. The precision for both categories reached 96.63%, which is critical for clinical applications to minimize false positives in pathological detection.
- **Crackle and Normal Cycles:** The detection of *Crackles* yielded an F1-score of 0.8848, while *Normal* respiratory cycles were identified with an F1-score of 0.8901. The high recall for the Normal class (0.9015) ensures that healthy respiratory cycles are correctly identified, reducing unnecessary clinical alarm.

The exceptional performance in the *Both* category is particularly significant, as this class represents the most complex acoustic signature in the dataset. The high metrics across all classes validate the effectiveness of the Triple-Stream architecture in fusing Mel, MFCC, and CQT features to capture the diverse spectral-temporal characteristics of respiratory sounds.

3.4 Training Stability

The learning dynamics of the Triple-Stream SelfONN model are illustrated in Fig. 3, which plots the training and validation accuracy over 250 epochs. A notable observation is the performance gap where the validation accuracy fluctuate between 92% and 93% consistently exceeds the training accuracy range (82% - 83%). While

Table 2. Detailed Classification Report for the 4-Class Task

Class	Precision	Recall	F1-Score
Normal	0.8790	0.9015	0.8901
Crackles	0.8806	0.8891	0.8848
Wheezes	0.9663	0.9624	0.9644
Both	0.9663	0.9362	0.9510

traditional learning curves typically show the inverse, this phenomenon is a direct result of the aggressive regularization and specialized loss functions implemented in the model framework.

Figure 3. Mean training and validation accuracy curves over 250 epochs using 10-fold stratified cross-validation. The solid lines denote the mean accuracy across folds, and the shaded regions represent ±1 standard deviation. This visualization provides a clearer summary of model convergence and inter-fold variability compared with plotting all individual fold curves. The higher validation accuracy may be related to the regularization and augmentation strategies applied during training, including Focal Loss and dropout.

The primary drivers for this gap are identified as follows:

- **Dropout Effect:** During training, dropout randomly disables parts of the network to reduce overfitting. During validation, dropout is turned off, allowing the full model to be used, which can improve validation performance.
- **Focal Loss Penalty:** Focal Loss makes the model focus more on difficult and minority classes, such as the "Both" class. This helps improve generalization instead of only learning the majority class.
- **Regularization:** L_2 weight decay was applied to prevent the model from memorizing the training data. This encourages the model to learn more general features that perform better on unseen samples.

As shown in Fig. 3, both curves converge and stabilize after approximately 200 epochs, indicating that the model has reached a state of optimal generalization without signs of divergence or traditional overfitting.

3.5 Comparison with Other Methods

To evaluate the effectiveness of the proposed Triple-Stream SelfONN, we compare our performance against established benchmarks on the ICBHI 2017 dataset.

As shown in Table 3, our model significantly outperforms early machine learning baselines and standard deep learning architectures. The SVM framework by Chambres et al. [7] illustrates the limitations of handcrafted features, while standard CNN and LSTM implementations [11, 12] demonstrate the plateau in accuracy (approx. 71–80%) when relying on single-feature inputs like STFT or Mel-spectrograms. Even specialized models like RespireNet [8], designed for limited data settings, and high-performing SincNet architectures [10] remain below the 91% threshold.

Our Triple-Stream approach achieves a superior accuracy of 92.23%, validating that the fusion of Mel, MFCC, and CQT features through Self-Operational layers provides a more robust representation of complex respiratory transients than traditional linear-convolutional pipelines.

Table 3. Comparison of Accuracy and F1-Score on the ICBHI 2017 Dataset

Model	Features	Acc	F1
SVM [7]	Handcrafted	49.98 %	-
CNN [11]	STFT	71.15%	65%
ResNet-34 [8]	Mel-Spec	71.81 %	70.20 %
CNN-LSTM [9]	STFT	76.39 %	68.52 %
LSTM [12]	Mel-Spec	79.61%	78.67%
SincNet [10]	Waveform	90.60 %	86.30 %
Proposed (Ours)	**PSSE**	**92.2%**	**92.3%**

4 Conclusion

This paper presented a novel framework for automated respiratory sound classification, explicitly designed to overcome the linear limitations of traditional CNN, SVM, and LSTM pipelines. By introducing the Parallel Spectral Stream Extraction (PSSE) module including Mel Spectrogram, MFCC, and Constant-Q Transform (CQT) and routing heterogeneous audio features through cubic Self-Operational Neural Networks (SelfONN), the system effectively isolates and classifies complex, non-linear pathological transients. Achieving 92% accuracy and F1-score on the ICBHI 2017 dataset, this architecture proves highly robust against class imbalance and ambient clinical noise. Future research will focus on deploying this highly efficient, shallow operational architecture onto edge-AI hardware for real-time intelligent stethoscope applications.

In future work, we plan to investigate the robustness of the system under challenging conditions, including ambient noise, heart sound interference, and variations in recording devices. Additionally, we aim to optimize the model for on-device deployment on embedded platforms, enabling fully portable and point-of-care diagnostic systems.

References

[1] World Health Organization, Chronic respiratory diseases, https://www.who.int/health-topics/chronic-respiratory-diseases (2023), accessed: 2024-04-16

[2] R.X.A. Pramono, S. Bowyer, E. Rodriguez-Villegas, Automatic adventitious respiratory sound analysis: A systematic review, PLOS ONE **12**, e0177926 (2017). 10.1371/journal.pone.0177926

[3] A.R.A. Sovijärvi, F. Dalmasso, J. Vanderschoot, L.P. Malmberg, G. Righini, M. Rossi, Standardization of lung sound nomenclature, European Respiratory Review **10**, 434 (2000).

[4] M. Sarkar, I. Madabhavi, N. Niranjan, M. Dogra, Auscultation of the respiratory system, Annals of Thoracic Medicine **10**, 158 (2015). 10.4103/1817-1737.160839

[5] A. Gurung, C.G. Scrafford, J.M. Tielsch, O.S. Levine, W. Checkley, Computerized lung sound analysis as diagnostic aid for the detection of abnormal lung sounds: A systematic review and meta-analysis, Respiratory Medicine **105**, 1396 (2011). 10.1016/j.rmed.2011.05.007

[6] E.D. McCollum et al., Digital auscultation in pediatric pneumonia: progress and opportunities, The Lancet Digital Health **1**, e421 (2019). 10.1016/S2589-7500(19)30154-1

[7] G. Chambres, P. Hanna, M. Desainte-Catherine, Automatic Detection of Patient with Respiratory Diseases Using Lung Sound Analysis, in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)* (2018), pp. 1–6, https://ieeexplore.ieee.org/document/8516481

[8] S. Gairola, F. Tom, N. Kwatra, 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)* (2021), pp. 527–530, https://ieeexplore.ieee.org/document/9630653

[9] G. Petmezas, G.A. Cheimariotis, L. Stefanopoulos, B. Rocha, R.P. Paiva, A.K. Katsaggelos, N. Maglaveras, Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function, Sensors **22**, 1232 (2022). 10.3390/s22031232

[10] S. Kulkarni, A. Guntoro, F. Al-Turjman, Self-Supervised Audio Encoder with Contrastive Pretraining for Respiratory Anomaly Detection, in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2023), pp. 1–5, https://ieeexplore.ieee.org/document/10095163

[11] F. Demir, A.M. Ismael, A. Sengur, Classification of lung sounds with CNN model using parallel pooling structure, IEEE Access **8**, 105376 (2020). 10.1109/ACCESS.2020.3000111

[12] R. Khan, S.U. Khan, U. Saeed, I.S. Koo, Auscultation-based pulmonary disease detection through parallel transformation and deep learning, Bioengineering **11**, 586 (2024). 10.3390/bioengineering11060586

[13] B.M. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimaris, P. Natsiavas, A. Oliveira, C. Jaccard, N. Martins et al., An Open Access Database for the Evaluation of Respiratory Sound Classification Algorithms, in *2017 IEEE Life Sciences Conference (LSC)* (2017), pp. 196–201, https://ieeexplore.ieee.org/document/8268165

[14] N. Jaitly, G.E. Hinton, Vocal Tract Length Perturbation (VTLP) Improves Speech Recognition, in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language* (2013), https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41183.pdf

[15] D.S. Park, W. Chan, Y. Zhang, C.C. Chiu, B. Zoph, E.D. Cubuk, Q.V. Le, SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, in *Proc. Interspeech 2019* (2019), pp. 2613–2617, https://www.isca-archive.org/interspeech_2019/park19_interspeech.html

[16] S.S. Stevens, J. Volkmann, E.B. Newman, A scale for the measurement of the psychological magnitude pitch, The Journal of the Acoustical Society of America **8**, 185 (1937). 10.1121/1.1915893

[17] S. Mondal et al., Mel-frequency spectral features for objective assessment of respiratory sounds, Biomedical Signal Processing and Control **71**, 103230 (2022). 10.1016/j.bspc.2021.103230

[18] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech, and Signal Processing **28**, 357 (1980). 10.1109/TASSP.1080.1163420

[19] M. Bahoura, Pattern classifier for the detection of wheezes and crackles, Computers in Biology and Medicine **39**, 25 (2009). 10.1016/j.compbiomed.2008.11.001

[20] J.C. Brown, Calculation of a constant Q spectral transform, The Journal of the Acoustical Society of America **89**, 425 (1991). 10.1121/1.400476

[21] J. Velayutham, S. Jayalakshmy, Multi-resolution analysis of respiratory sounds for pulmonary disorder detection, IEEE Sensors Journal **23**, 17351 (2023). 10.1109/JSEN.2023.3283626

[22] A. Roy, U. Satija, AsTFSONN: A Unified Framework Based on Time-Frequency Domain Self-Operational Neural Network for Asthmatic Lung Sound Classification, in *2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (2023), pp. 1–6, https://ieeexplore.ieee.org/document/10171911

[23] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2980–2988, https://ieeexplore.ieee.org/document/8237586