Supervised Contrastive Pretrained ResNet with MixUp to Enhance Respiratory Sound Classification on Imbalanced and Limited Dataset

Jinhai Hu[†*], Cong Sheng Leow[*], Shuailin Tao[†*], Wang Ling Goh[†], Yuan Gao[*]
†School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Republic of Singapore
*Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), Republic of Singapore
Email: jinhai001@e.ntu.edu.sg, leow_cong_sheng@ime.a-star.edu.sg, shuailin001@e.ntu.edu.sg, ewlgoh@ntu.edu.sg, gaoy@ime.a-star.edu.sg

***Abstract*—This paper proposes a strategy of combining multiple techniques to classify paediatric respiratory sound (PRS) from the Open-Source SJTU Paediatric Respiratory Sound Database. Inspired by recent successes in image classification, this work focuses on improving audio classification with limited and imbalanced datasets through Residual Networks (ResNet). These techniques include augmentations applied to audio features, supervised contrastive (SupCon) pretraining, and MixUp. These three techniques helped reduced overfitting due to imbalanced dataset. To further enhance accuracy, pre-processing, and training hyperparameters were optimized through Bayesian Optimization. The proposed strategy achieved over 95% training accuracies for the four tasks (11, 12, 21, and 22) in the IEEE BioCAS 2023 grand challenge. Through this strategy, the four tasks achieved calculated scores of 0.769, 0.632, 0.662 and 0.512 respectively using the test dataset. The total score is 0.729 including 0.1 obtained from the runtime bonus.**

Keywords—Supervised contrastive learning, MixUp finetuning, Respiratory sound classification

I. Introduction

With one of the highest mortality rates, early detection of respiratory diseases is pivotal in both the treatment and the maintenance of health conditions [1]. Auscultation can be used to identify abnormal respiratory sounds for diagnosis but is limited in its effectiveness. Auscultation requires experts with relevant experience, and even then, variability exists across multiple experts due to varying individual auditory systems [2]. Furthermore, intermittent abnormalities in some of these respiratory diseases make continuous monitoring required but challenging for conventional auscultation approaches [1].

To address the accessibility, reliability, and consistency of respiratory disease diagnosis, many studies on automated respiratory sound classification systems have been done as listed in [3], [4]. From the 2022 BioCAS Grand Challenge [4], Convolutional neural networks (CNN) and ResNets were popular among the top few winning teams. By converting to frequency-based features such as Mel-Cepstral Frequency Coefficients (MFCCs) or Short-time Fourier Transform (STFT), CNNs can extract key feature representation through the deep layers. For instance, [5] and [6] proposed ResNet while [6] and [7] proposed CNN.

However, one pertinent issue with the existing respiratory sound databases is the class imbalance problem, where the normal audio heavily dominates the dataset. This imbalance leads to biases and low recall. To circumvent this problem, data augmentation techniques have been commonly employed. In last year's grand challenge, [5], [6], [8] proposed samplers and different loss functions to reduce class imbalance. Nonetheless, current work is still lacking in imbalanced datasets, which better represents the real-world data.

Building on previous works, this study seeks to improve the effectiveness of paediatric respiratory sound classification. With ResNet being highly successful such as in [5], we similarly propose a ResNet architecture. In addition to exploring various preprocessing methods, we focused on alleviating the impact of limited and imbalanced datasets in this work. Originally designed for vision tasks with limited data [9], the SupCon approach improved classification capability by being more sensitive to the subtle variations in the MFCC for the respiratory sound of different anomalies. Furthermore, the MixUp method showed potential in previous studies for imbalanced data [10], [11], and we saw similar positive outcomes with respiratory audio data [12]. The contributions of this work are summed up as followed:

- We illustrate the use of image augmentation techniques for MFCC augmentation, namely random crop, and random flip in the preprocessing step.
- We implemented SupCon to improve the performance of ResNet using a limited respiratory sound dataset.
- We implemented MixUp to further enhance classification ability with the imbalanced dataset.

The strategy achieved competitive performance in the four tasks listed under the IEEE BioCAS Grand Challenge [4]. To our best knowledge, there has not been any work exploring the combination of the three techniques in addressing imbalanced respiratory audio dataset. This paper is organized as follows: Section II introduces the proposed models and techniques. Section III discusses the results of PRS classification and benchmarks against comparable works. Section IV summarises and concludes the work.

II. Proposed Models

The block diagram in Fig. 1 illustrates the full flow of our proposed strategy, including the SupCon pretraining and MixUp finetuning steps. First, the raw PRS signal was preprocessed to extract MFCC feature with augmentation techniques. The SupCon pretrained model encoded the feature into a high-dimensional embedding vector where data with the same label are close together while the distance between different labels are further apart. The linear classifier was used for classification and where finetuning can be done. The MixUp method was employed during finetuning to mitigate

The first two authors contributed equally to this work. This work was supported by the Agency for Science, Technology and Research (A*STAR), Singapore under the Cyber-Physiochemical Interface programme, grant No. A18A1b0045 and the Nanosystems at the Edge programme, grant No. A18A1b0055.

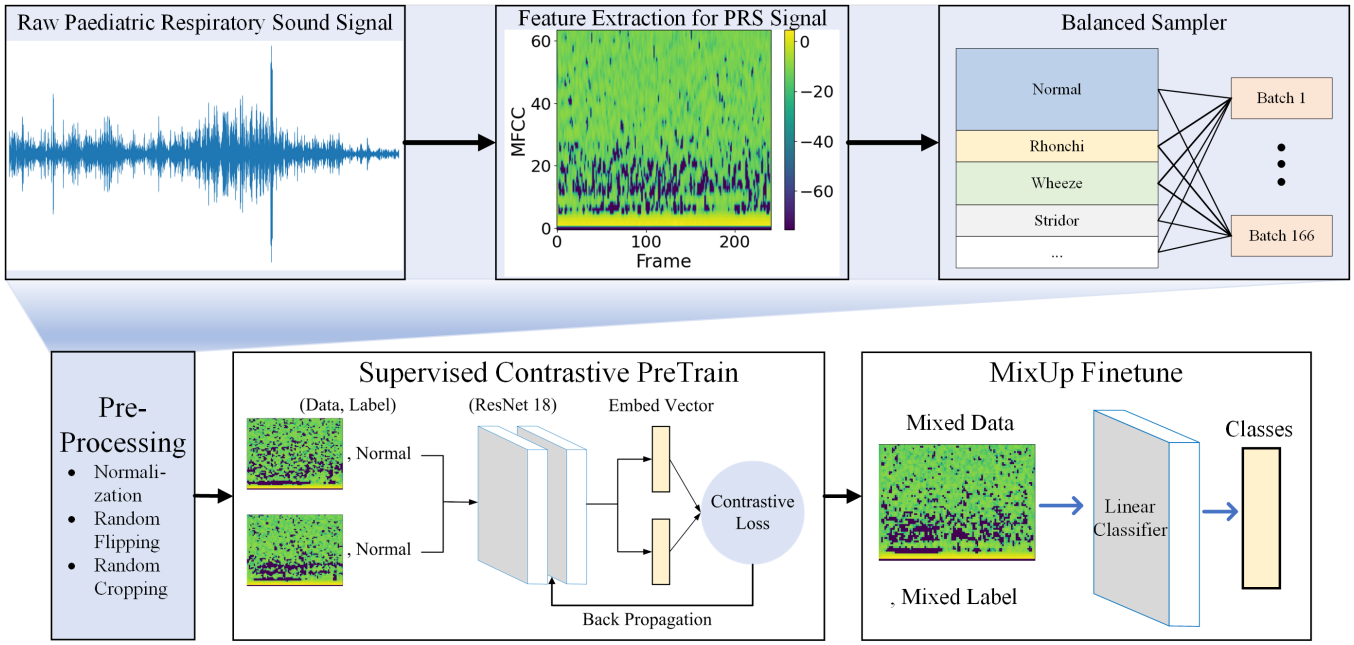


Fig. 1. Schematic diagram for overall model, which includes data pre-processing, supervised contrastive pretraining and MixUp finetuning.

overfitting caused by the limited training data. By specifying different tasks, a different configuration can be used to configure the model and the hyperparameters. For the classification of the entire recording (record level) instead of clipped audio (event level), we attempted real-time inference using a sliding-window approach. This is further explained in Section III.C.

A. Data Processing and Balanced Sampler

Pre-normalization in the time domain was used to scale each signal to an amplitude of ±0.5 and a mean of 0. Following this, MFCC was extracted with optimized hop length and number of filter channels (elaborated further in Section III.B). To ensure equalized input levels when feeding the data into the cascaded SupCon pretrain model, Z-score normalization was performed on the feature.

To reduce biases due to an imbalanced dataset, a balanced sampler was used in the construction of DataLoaders. For balanced samplers, a weight is assigned to each class reciprocal to the amount of sample present in the dataset. The weighted random sampler ensures that every batch includes a similar number of samples from each category by randomly selecting samples from each class based on the weights. Moreover, to prevent the selection of identical samples through repeated sampling, various transforms, such as random crop and random flip, were implemented in the collate function during the generation of DataLoaders. These transforms introduce diversity and variation into the samples within each batch.

Each MFCC initially had the same number of channels, but a different number of elements along the time axis. Random cropping provided two-fold advantages in this context. Firstly, it allowed us to crop every sample to a consistent length for feeding into the subsequent neural network. For samples that exceeded the desired window size, random cropping trimmed them to a shorter length randomly. Similarly, for shorter samples were randomly padded with zeroes, represented by the mean value after z-score normalization. The second advantage of random cropping is that it increased the robustness of the neural network. Random cropping simulated event shift along the temporal axis, enabling the network to develop increased tolerance to temporal variations and shifts within the PRS signals.

Despite the variation in the order between exhale and inhale signals, the label for each breath should remain the same. To enable the model to effectively recognize this bidirectional information, random flipping was applied as a data augmentation technique. Random flipping allowed the model to learn and generalize better to the different directional patterns present in the PRS data. This enhanced the model's ability to accurately classify respiratory sounds regardless of the breathing cycles.

However, to ensure the integrity and accuracy of the test results, the DataLoader for the test dataset retained the default sampler without any repetition or shuffling. This ensures that each sample in the test set is presented in its original order, without any randomization that may introduce bias or affect the evaluation of the model's performance. By maintaining the default sampler, we could obtain reliable and unbiased test results for the PRS classification model.

B. Supervised Contrastive Pretraining

In recent years, contrastive learning [13] has emerged as a powerful technique in the field of representation learning. While contrastive learning initially gained prominence in self-supervised settings, it has shown remarkable success in unsupervised training of deep image models [14]–[16] and even learning sound events [17]. Modern batch contrastive approaches have not only surpassed but also significantly outperformed traditional contrastive losses. [9] extended the application of contrastive learning from self-supervised to the fully-supervised setting, enabling us to effectively leverage label information when available. By applying SupCon pretraining, the pretrained model became adept at learning robust and discriminative representations, which are crucial for accurate PRS classification.

The SupCon DataLoader was created using two crop transformations, where each feature map passed through a composed data augmentation process. This augmentation aims to increase the variability within signals sharing the same label. This pair of contrastive data passes through the pretrain model during each iteration of pretraining to give embedding vectors

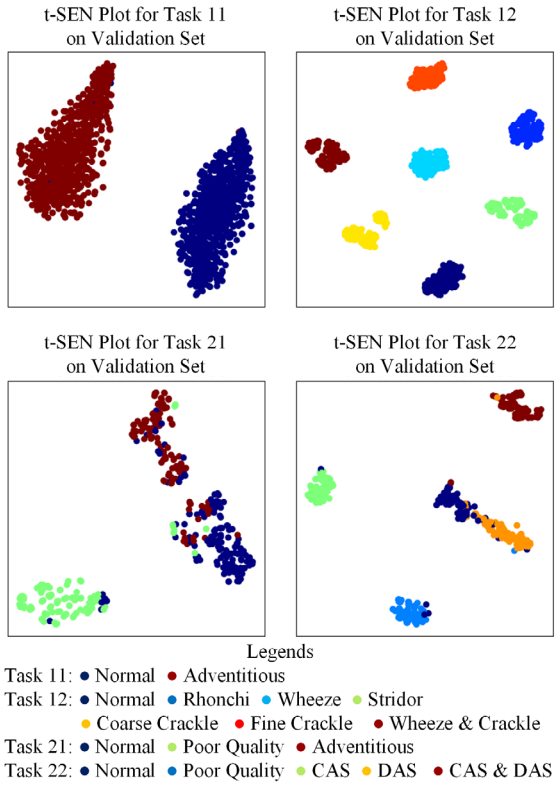


Fig. 2. t-SEN Plot for Task 11, 12, 21 and 22 on validation datasets.

as the output, where contrastive loss can be calculated. This loss was then used for updating the weights of the pretrained model. In the pretrained model, audio samples with the same label were clustered together in the embedding space, while simultaneously maximizing the distance between different labels. The representation of each signal was captured as an embedded array, and the projection diagram illustrating these embeddings can be visualized using t-SNE plots, as shown in Fig. 2.

III. Optimization

To further improve the performance of the proposed network, optimization was done to further address data scarcity and to choose the right parameters in preprocessing and training.

A. MixUp Finetuning

SupCon pretraining played a crucial role in significantly reducing the initial loss and improving the accuracy during the subsequent fine-tuning process, which involves the classification using a linear classifier. Furthermore, a recent paper has also confirmed the success of SupCon for respiratory audio [18]. However, if the same dataset is used for both pretraining and finetuning, there is a risk of overfitting, particularly when working with limited trainable data. To address this issue, the MixUp method was used. The MixUp method was initially introduced to regularize the network and demonstrated its efficacy in enhancing the performance of classification models [11], [12]. The regularization effect helped in preventing overfitting, especially with a limited dataset. The MixUp method involves creating new samples by combining original samples with others, wherein the labels for the mixed samples are also generated based on their constituent labels. The MixUp process is represented as follows [11]:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \quad (1)$$

$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \quad (2)$$

Table I Comparison on Performance of MixUp Method

Topology	Dataset	Task 11	Task 12	Task 21	Task 2-2
Linear Classifier	Train	0.998	0.976	0.938	0.895
	Valid	0.898	0.854	0.785	0.708
	Inter-Test	0.836	0.801	0.731	0.557
+ MixUp	Train	0.950	0.924	0.874	0.826
	Valid	0.913	0.866	0.792	0.711
	Inter-Test	0.851	0.816	0.745	0.579

Here, λ is controlled by the MixUp interpolation coefficient α and follows a beta distribution $X \sim Beta(\alpha, \alpha)$. By training the model on these synthesized samples, the MixUp method encourages the model to learn from the relationships and patterns present in different samples. Although the MixUp method may lead to a slight degradation in training accuracy, it showed improvements in both validation and test scores during the finetuning process for our implementation. The specific results can be observed in Table I. We focused on inter-test scores as the due to increased variability across patients.

B. Tuning of Preprocessing and Hyperparameters

We employed Bayesian optimization to further improve our model using *Tune* [19]. This optimization process involved tuning both the hyperparameters and the preprocessing parameters to maximize performance. For the hyperparameters, the search space included important factors such as the learning rate (1e-4, 1e-1) the number of convolution layer filters, and dense layers (in power of two's). In the case of supervised contrastive training, the main parameters considered in the search space were the temperature (0.1, 0.9) and learning rate.

Preprocessing parameters including the hop length, the number of MFCC coefficients, and the FFT length, were also optimized to enhance the preprocessing stage. Eventually, the MFCC features were extracted using the optimal parameters. This systematic tuning of both network and preprocessing hyperparameters allowed us to identify the optimal configuration that maximized the performance of our model for PRS classification.

C. Real-time Inference for Record level Task

For task 2, the classification of PRS signals can be performed in real-time by implementing a sliding-window approach instead of classifying with the full audio recording. The label assigned at the record level is closely related to the event level classification within each record.

The events can be categorized into Continuous Adventitious Sounds (CAS), and Discontinuous Adventitious Sounds (DAS). CAS includes include Rhonchi, Wheeze, and Stridor, while DAS includes Coarse Crackle and Fine Crackle. If any of these adventitious events are detected within a PRS record, the record is classified accordingly as CAS or DAS. Conversely, if no normal or adventitious events are identified in the entire record, it is classified as a "Poor Quality" record. By applying the sliding-window technique, real-time inference can be achieved by continuously analyzing the signal within the sliding window. This allows for the monitoring of the lung sound status in real-time, providing event outputs with minimal latency. Additionally, the occurrence of event level labels can be used to further cluster and analyze the record level labels, providing valuable insights into the distribution and occurrence patterns of different events within the PRS recordings.

To implement the proposed algorithm, we retrained the event level model by extending the event shape to match the

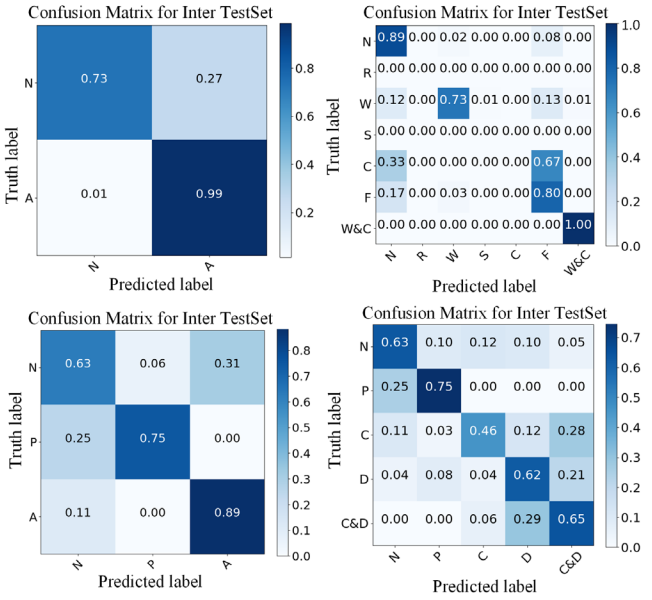


Fig. 3. Confusion Matrix for Task 11, 12, 21 and 22 on inter-test datasets.

window size. Additionally, we introduced a new category for poor-quality events. When a record is labelled as poor quality, the isolation function selects a poor-quality event using random cropping to reduce repetition within batches. Considering the characteristics of different events [1], where most events have a duration of fewer than three seconds, we set a window length of three with a one-second overlap between consecutive windows. This design choice helped prevent accidental identifications and increased the model's tolerance for event recognition. Specifically, for a positive classification at record level, two consecutive repeated events must be detected. If abnormal events are present in the record, the record is classified as an adventitious event. Furthermore, we checked for the presence of normal events. If there are only poor-quality events in a record, it is classified as poor quality.

IV. Results and Discussion

A. Classification Results

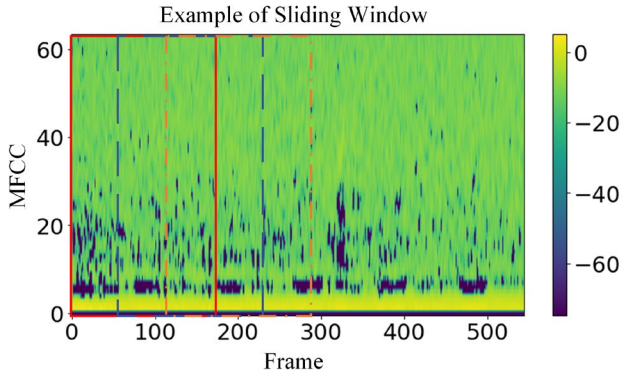
The performances of all four tasks (without real-time inference) are shown in the confusion matrix in Fig. 3, with the scores calculated based on [4] shown in following subsections. Some of the events have zero samples in provided inter-test dataset, leading to rows of zeros.

B. Real-time Inference Challenges

Fig. 4 shows an example of the sliding-window and the classification for each window, whose window size is three seconds, and the corresponding sliding length is one second. The conversion between the period and number of frames in the feature map can be found in equation (3), where the sampling rate is 8000 and the hop length is 128.

$$Frame = \frac{time(s) \times Sampling\ Rate\ (Hz)}{hop\ length} \quad (3)$$

While the sliding-window approach incur delay in overall computation, it offers real-time classification and accurately indicates the exact occurrence time of adventitious events as in Fig. 4. The average score obtained for task 21 and 22 for both inter- and intra-classification are 0.835 and 0.680 respectively. However, the decision-making process to integrate the event level classification into a record level classification remains challenging. The average final record level score of 0.539 And 0.436 are obtained for task 21 and task 22 respectively as shown in Table II. In this study, the



Sliding Window (s)	0-3	1-4	2-5	3-6	4-7	5-8
Predict Label	Fine Crackle	Fine Crackle	Fine Crackle	Wheeze	Wheeze	Wheeze
Event Period (s)		1.2-1.8	1.8-2.4	3.0-3.6	4.1-4.6	
Truth Label		Fine Crackle	Wheeze	Wheeze	Wheeze	

Fig. 4. Example of sliding-window based PRS signal recognition.

TABLE II Sliding Window Scores

Classification Level	Task 21 (intra)	Task 22 (intra)	Task 21 (inter)	Task 22 (inter)
Event level	0.840	0.660	0.830	0.700
Record level (integrated)	0.548	0.431	0.530	0.440

TABLE III Benchmark Table

Topology	Task 11	Task 12	Task 21	Task 22
CNN	0.728	0.526	0.537	0.413
+ MixUp	0.744	0.552	0.531	0.415
ResNet 18	0.708	0.576	0.557	0.443
+ SupCon	0.754	0.617	0.648	0.490
+ SupCon + MixUp	**0.769**	0.625	0.658	0.494
+ SupCon + MixUp + Tune	0.762	**0.632**	**0.662**	**0.512**

naive algorithm to check for consecutive classifications may fail for real-life implementation due to variability of the anomalies in respiratory sound. As such, we decided that our sliding-window approach is not mature enough for PRS classification and reverted to stand-alone record level classification for task 2, with the inter-test results shown in Fig. 3 and final test results shown in Table III.

C. Benchmarking and Discussion

As seen in Table I, the training accuracy decreased with the use of MixUp while test accuracies increased, confirming the problem of overfitting. Table III further shows how the different techniques contributed to the improvement in performance (final test score).

SupCon and MixUp simulated effects of having a larger dataset, making the proposed approach well-suited for constructing a pretraining model with high specificity and precision. To address the issue of limited data, we enhanced the algorithm by incorporating MixUp into the supervised contrastive loss. This approach effectively tackles the problem of data insufficiency during the pretraining process.

V. Conclusion

In this paper, a comprehensive approach to PRS classification is introduced. Using augmentation, SupCon and MixUp, we were able to improve the network's robustness against overfitting due to a limited dataset. The sliding-window approach provided valuable information on the occurrence of adventitious events and enabled real-time classification. However, real-time classification remains challenging. Furthermore, developments on decision-making for the sliding-window approach will improve the real-time classification performance. Access to larger datasets will also contribute to refining and expanding the applicability of our approach in clinical settings.

REFERENCES

[1] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," PLOS ONE, vol. 12, no. 5, May 2017.

[2] H. Kiyokawa, M. Greenberg, K. Shirota, and H. Pasterkamp, "Auditory Detection of Simulated Crackles in Breath Sounds," *Chest*, vol. 119, no. 6, pp. 1886–1892, Jun. 2001.

[3] Q. Zhang *et al*., "SPRSound: Open-Source SJTU Paediatric Respiratory Sound Database," *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 5, pp. 867–881, Oct. 2022.

[4] Q. Zhang *et al*., "Grand Challenge on Respiratory Sound Classification for SPRSound Dataset," *IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, pp. 213–217, Oct. 2022..

[5] J. Li, X. Wang, X. Wang, S. Qiao, and Y. Zhou, "Improving The ResNet-based Respiratory Sound Classification Systems With Focal Loss," *IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, pp. 223–227, Oct. 2022.

[6] Z. Chen, H. Wang, C.-H. Yeh, and X. Liu, "Classify Respiratory Abnormality in Lung Sounds Using STFT and a Fine-Tuned ResNet18 Network," *IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, pp. 233–237, Oct. 2022.

[7] N. Babu, J. Kumari, J. Mathew, U. Satija, and A. Mondal, "Multiclass Categorisation of Respiratory Sound Signals Using Neural Network," *IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, pp. 228–232, Oct. 2022.

[8] L. Zhang, Y. Zhu, S. Tu, and L. Xu, "A Feature Polymerized Based Two-Level Ensemble Model for Respiratory Sound Classification," *IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, pp. 238–242, Oct. 2022.

[9] P. Khosla *et al*., "Supervised Contrastive Learning," *Advances Neural Inf. Process. Syst. (NIPS)*, pp. 18661–18673, 2020.

[10] A. Galdran, G. Carneiro, and M. A. González Ballester, "Balanced-MixUp for Highly Imbalanced Medical Image Classification," *Med. Image Comput. Comput. Assisted Intervention (MICCAI)*, 2021.

[11] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," arXiv, Apr. 2018. doi: 10.48550/arXiv.1710.09412.

[12] Y. Ma, X. Xu, and Y. Li, "LungRN+NL: An Improved Adventitious Lung Sound Classification Using Non-Local Block ResNet Neural Network with Mixup Data Augmentation," in *Proc. Interspeech*, pp. 2902–2906, Nov. 2020.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," arXiv, Jun. 2020. doi: 10.48550/arXiv.2002.05709.

[14] E. Xie *et al.*, "DetCo: Unsupervised Contrastive Learning for Object Detection," *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 8392–8401, 2021.

[15] Z. Wen and Y. Li, "Toward Understanding the Feature Learning Process of Self-supervised Contrastive Learning," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, pp. 11112–11122, Jul. 2021.

[16] X. Liu *et al.*, "Self-Supervised Learning: Generative or Contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, Jan. 2023, doi: 10.1109/TKDE.2021.3090866.

[17] E. Fonseca, D. Ortego, K. McGuinness, N. E. O'Connor, and X. Serra, "Unsupervised Contrastive Learning of Sound Event Representations," *IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, pp. 371–375, Jun. 2021.

[18] I. Moummad and N. Farrugia, "Learning Audio Features with Metadata and Contrastive Learning," arXiv, Mar. 2023. doi: 10.48550/arXiv.2210.16192.

[19] P. Moritz *et al.*, "Ray: A Distributed Framework for Emerging AI Applications," *USENIX Symp. on Operating Syst. Des. Implementation (OSDI 18)*, pp. 561–577, 2018.