ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447967

# GAP-AUG: GAMMA PATCH-WISE CORRECTION AUGMENTATION METHOD FOR RESPIRATORY SOUND CLASSIFICATION

*An-Yan Chang*[1], *Jing-Tong Tzeng*[2], *Huan-Yu Chen*[1], *Chih-Wei Sung*[3,4], *Chun-Hsiang Huang*[4], *Edward Pei-Chuan Huang*[3,4], *Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]College of Semiconductor Research, National Tsing Hua University, Taiwan
[3]Department of Emergency Medicine, College of Medicine, National Taiwan University, Taiwan
[4]Department of Emergency Medicine, National Taiwan University Hsin-Chu Hospital, Taiwan

## ABSTRACT

Automated auscultation analysis using electronic stethoscope has received growing interest in clinical applications. Recently, researchers showed successes by using deep learning methods to distinguish between pathological respiratory sound classes. Nevertheless, the challenge persists due to the scarcity of abnormal samples, and the distinct characteristics between low-pitched and discontinuous *crackles* and high-pitched and continuous *wheezes*. In this study, we proposed a novel augmentation method, namely gamma patch-wise correction augmentation, which directly operates on spectrograms to handle with these two challenges. We achieved state-of-the-art performances on both 60-40 official split and 80-20 cross-validation of the public ICBHI dataset, outperforming previous top-performing studies by 11.82% in sensitivity and 5.27% in ICBHI score. Furthermore, Grad-CAM analysis shows that our approach better preserves the distinctive characteristics of *crackles* and *wheezes* than SpecAug.

***Index Terms***— Data augmentation, Respiratory sound classification, Gamma correction, Mix up

## 1. INTRODUCTION

Respiratory system diseases represent a leading cause of mortality worldwide [1], imposing a substantial burden on a global scale. Under the epidemic of COVID-19, the demand of non-invasive and rapid diagnosis is increasing. The evolution of stethoscope and advancement of computational methods open up opportunities for automated examination of respiratory sounds, reducing the demand for manual efforts from physicians and medical professionals. A notable milestone for the field is the release of the International Conference on Biomedical and Health Informatics (ICBHI) dataset [2]. It is a publicly accessible repository, designed to investigate respiratory pathology using stethoscope, i.e., *normal* respiration, *crackles*, and *wheezes*. Recently, works that take spectrograms as input for deep networks have achieved state-of-the-art (SOTA) classification performances on these auscultatory sound recordings (e.g., using structures of convolutional and recurrent neural networks [3–5]).

Many of the past works show a high score on identifying normal respiration (most common class) but only a moderate performance on detecting abnormal sounds [6–8]. Reliable identification of abnormal classes, including *crackles* and *wheezes*, is pivotal for precise diagnosis and remains a major computational challenge. *Crackles* and *wheezes* have distinct characteristics. *Crackles* are discontinuous and low-pitched with non-musical auditory patterns linked to conditions like congestive heart failure, pneumonia, and lung fibrosis; *wheezes* are continuous high-pitched tonal sequences observed in individuals with asthma, COPD, or tumors [9]. An advancing method to handle the limited availability of pathological samples and the subtle distinct characteristics that differentiate between the two abnormal classes is key to improve performances. Instead of many recent works that delved into designing heavily complex network architectures for respiratory sound classification [7, 8, 10], we propose to handle these issues with simple yet effective data augmentation strategy.

In spectrogram-based audio and speech modeling tasks, SpecAugment (SpecAug) [11] has become the de-facto method of feature augmentation. We argue that SpecAug, i.e., randomly masking strips or patches of spectrogram, might eliminate those high or low-frequency bands that contain critical acoustic signatures, especially for abnormal breathing classes (Fig 1). In this work, we propose a novel augmentation process, gamma patch-wise correction augmentation (GaP-aug), that operates on log mel spectrogram. The augmented dataset are generated with the mixup strategy [12], where each to-be-mixed spectrogram first goes through a random patch-wise masking with gamma correction. GaP-aug preserves the continuous presence of *wheezes* and the explosive burst of *crackles* by adjusting contrast within patches. Our method improves the leading study by 5.27% in ICBHI score, and further analysis reveals our method better preserves key distinctive structures of *crackles* and *wheezes*.

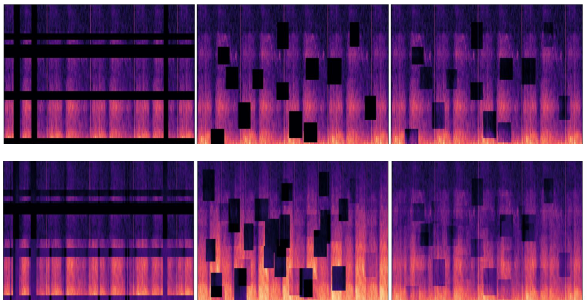


**Fig. 1**: (left) SpecAug; (mid) Patch masking; (right) GaP-aug. The bottom row is with mixup

## 2. METHODOLOGY

### 2.1. Dataset

ICBHI 2017 dataset is composed of respiratory recordings collected from 126 participants using a number of stethoscopes, comprising 6898 cycles with approximately 5.5 hours in total. It includes four sound classes: *Both* (crackle and wheeze), *normal*, *crackle* and *wheeze*. The label distribution is heavily imbalanced: 3642 normal respiratory cycles, 1864 crackles, 886 wheezes, and 506 both crackles and wheezes. We use the evaluation metrics proposed in the ICBHI challenge [2], including sensitivity (Sen.), which is recall of abnormal; specificity (Spe.), which is recall of normal; and final ICBHI score, the average of sensitivity and specificity.

### 2.2. Gamma patch-wise correction augmentation

The proposed GaP-aug operates by generating augmented instances using mixup technique applied to patch-wise gamma corrected spectrograms.

#### *2.2.1. Mixup*

The Mixup method [12] is a widely adopted data-agnostic technique applicable to various modalities to construct synthetic training samples and has found its success in deep learning-based applications, e.g., Patch-Mix [6], TPH-YOLOv5 [13]. The process works by combining two training samples with their respective labels to generate a new augmented sample, described as follows:

$$\begin{aligned} \tilde{x} &= \lambda \tilde{x_i} + (1-\lambda)\tilde{x_j} \\ y &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{1}$$

where $\tilde{x_i}$, $\tilde{x_j}$ are features (spectrograms corrected by GaP-aug in this case), and $y_i$, $y_j$ are their one-hot labels. $\lambda \in [0, 1]$ is a randomly generated value following the Beta distribution, and it represents the probability of the new produced label vector of each class. The mixup method facilitates the linear transition of decision boundaries between classes and contributes to a smoother estimation of uncertainty.

**Algorithm 1** Pipeline of a batch

**Parameters** $B$ = batch size; $\mathcal{M}$ = model; $\mathcal{L}$ = loss
**Input:** $X = \{x_1, x_2, ...x_B\}$; $Y = \{y_1, y_2, ...y_B\}$

**procedure** *GaP-aug*($x_i$) ▷ a spectrogram
  Select random patches $P$ within $x_i$
  **for** $p_i \in P$ **do**
    $\tilde{p_i} \leftarrow Gain(p_i)^{\gamma}$
**for** $x_i \in X$ **do** ▷ $\tilde{X}_i \leftarrow X_i$
  $\tilde{x_i} \leftarrow$ *GaP-aug*($x_i$)
$\tilde{X}_a \leftarrow \{\tilde{x}_1, \tilde{x}_3, ..., \tilde{x}_{B-1}\}$; $Y_a \leftarrow \{y_1, y_3, ..., y_{B-1}\}$
$\tilde{X}_b \leftarrow \{\tilde{x}_2, \tilde{x}_4, ..., \tilde{x}_B\}$; $Y_b \leftarrow \{y_2, y_4, ..., y_B\}$
**for** $i = 1$ *to* $\frac{B}{2}$ **do** ▷ mixup
  Random generate $\lambda \in [0, 1]$
  $\tilde{X}_{mix}(i) \leftarrow \lambda \tilde{X}_a(i) + (1-\lambda)\tilde{X}_b(i)$
  $Y_{mix}(i) \leftarrow \lambda Y_a(i) + (1-\lambda) Y_b(i)$
*prediction* $\leftarrow \mathcal{M}(\tilde{X}_{mix})$; *embedding* $\leftarrow \mathcal{M}(\tilde{X}_a)$
$\mathcal{L} \leftarrow \mathcal{L}_{NLL}$(*prediction*, $Y_{mix}$) + $\mathcal{L}_{triplet}$(*embedding*)

#### *2.2.2. Gamma patch-wise correction*

Gamma correction is a technique mostly found in image processing that involves using a nonlinear power-law transformation to each pixel, consequently rescaling the contrast of input images. The formula is shown in Eqn. 2, where $p$ and $\tilde{p}$ indicates input and output selected patches, and gain is typically set as 1.

$$\tilde{p} = Gain(p)^{\gamma} \tag{2}$$

The variable $\gamma$ affects the brightness of chosen pixels, and when set to a high value, induces augmentation effects similar to noise suppression [14]. Based on grid search, $\gamma$ is uniformly randomized within the interval [1.7, 2.0] in this work, emphasizing strong and suppressing weak signals. Furthermore, patches of unfixed sizes are randomly selected from the spectrogram during enumeration. Through grid search, we select 32 patches per spectrogram where each is no larger than 16 in frequency multiplied by 16 in time. A batch-wise GaP-aug is shown in Algo. 1, and examples of mixup augmentation of SpecAug, patch masking, and GaP-aug are shown in Fig 1.

### 2.3. Respiratory sound classifier

We use AudioSet pretrained CNN14, introduced in [18], as backbone of our model that takes input of log mel spectrogram. This pre-trained model provides a high capacity representation power alleviating issues of training from scratch where large scale data availability is often an issue in medical domain. After the GaP-aug process, we then fine-tune on the augmented set using NLL loss and apply an additional triplet constraint on the embedding to further encourage class-wise discriminability. This fine-tuned CNN14 after GaP-aug serves as our respiratory sound classifier.

**Table 1**: Performance comparison with SOTA. Underline denotes the previous state-of-the-art, and **Bold** denotes best scores. Audio denotes speed, loudness and time shift adjustment. Concat denotes concatenation augmentation. Clip denotes blank clipping. Domain denotes domain transfer. Overlap denotes window overlapping.

| Split | Method | Architecture | Aug | Acc.(%) | Sen.(%) | Spe.(%) | **Score**(%) |
|---|---|---|---|---|---|---|---|
| 60-40 | Cotuning [8] | ResNet | – | – | 37.24 | 79.34 | 58.29 |
| | RespireNet [15] | ResNet34 | Concat, Clip | – | 40.10 | 72.30 | 56.20 |
| | Domain Transfer [16] | ResNeSt | Domain | – | 40.20 | 70.40 | 55.30 |
| | ARSC-Net [10] | bi-ResNet-Att | Audio, Mixup | – | 46.38 | 67.13 | 56.76 |
| | Metadata [3] | CNN6 | SpecAug | – | 39.15 | 75.95 | 57.55 |
| | Patch-Mix CL [6] | AST | Patch-Mix | – | 43.07 | **81.66** | 62.37 |
| | **Ours** | CNN14 | **GaP-aug**, Mixup | 69.01 | **58.20** | 77.07 | **67.64** |
| 80-20 | RespireNet [15] | ResNet34 | Concat, Clip | – | 53.70 | 83.30 | 68.50 |
| | LSTM-S7 [5] | RNN | Overlap | – | 62.00 | 85.00 | 74.00 |
| | MBTCNSE [7] | TCN | Overlap | 72.50 | 65.30 | **86.10** | 75.70 |
| | Multi-feature [17] | CNN | Audio | – | 67.22 | 82.87 | 75.04 |
| | Contrastive Embed [4] | CNN | Audio | 78.73 | 70.93 | 85.44 | 78.18 |
| | AudioSet pretrained [18] | CNN | – | 64.80 | 43.38 | 83.93 | 63.66 |
| | **Ours** | CNN14 | **GaP-aug**, Mixup | **80.70** | **74.62** | **86.13** | **80.37** |

## 3. EXPERIMENTAL SETUP AND RESULTS

### 3.1. Experiment setup

We evaluate on the official split of ICBHI, where patients are divided into training (60%) and testing (40%) set without overlap. We also report results on 5-fold cross-validation (80%-20% split) following many recent works on this dataset [4, 5, 7, 15, 17]. The sample rate of the dataset varies from 4kHz to 44.1kHz, we resample all recordings to 16kHz as previous works [3, 4, 6]. Then, each respiratory cycle is divided into 10-second audio segments, and we concatenate shorter cycles to 10 seconds. Next, the audio waveform is converted into 64-dimensional log Mel filterbank with 25ms window length and 6.25ms hop size to generate spectrograms. For the CNN model, we train 20000 iterations using Adam optimizer, 1e-5 learning rate and batch size of 16. Instances are balanced sampled in each learning batch, i.e., that same number of samples from each class are loaded in a batch.

### 3.2. Classification results

Table 1 summarizes the classification performances of our approach and other classification methods in the literature on the ICBHI dataset. We have achieved an increase of 11.82% in sensitivity and 5.27% in final score compared to the top-performing studies on 60-40 split dataset (sen. [10], score [6]). On the 80-20 split of 5-fold cross-validation, we outperform the leading study [4] with improvements of 3.69% in sensitivity and 2.19% in final score. A notable observation is that our approach shows its strength in improving sensitivity while prior studies often struggle. The ability of our GaP-aug to positively identify abnormal breathing is key to the improvement. In fact, by comparing 'Ours' to 'AudioSet pretrained' (the same backbone pretrained CNN model without GaP-aug), we see that there is a 31.24% improvement of sensitivity after applying our proposed GaP-aug. Given the significance of early diagnosis and treatment, our method improves on overall performances and notably on sensitivity, which is important particularly for clinical applications.

#### *3.2.1. Comparison to other augmentation methods*

We further compare performances against other methods of augmentations using the same backbone model architecture. The compared methods include: direct augmentations on audio, such as white noise addition, speed alternation, loudness adjustment and time-shift manipulation; concatenate augmentation, introduced by [15], works by concatenating abnormal waveforms as new instances; mixup metric learning [19] applies mixup method on embeddings, employing soft-label weights to separate embeddings of distinct classes with multi-similarity loss; finally, the rest are spectrogram-based, blank clipping [15] clips out insignificant segments, SpecAug [11] masks specific time and frequency intervals, and PatchMask randomly masks patches within spectrograms.

Table 2 summarizes our performance on the official 60-40 split when compared to the above-mentioned audio augmentation techniques. One thing to note that almost all augmentation methods help improve the classification performances, and our proposed GaP-aug improves the most, i.e., 8.15% on sensitivity and 10.41% on ICBHI score, compared to method without augmentation, referred as the 'Naïve' approach. Mixup strategy further enhances the performances by comparing Gap-aug w/o Mixup and Gap-aug w/ Mixup.

**Table 2**: Performance comparison of augmentations.

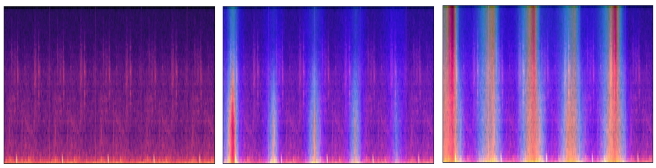
| Augmentations | base on | **Sen.**(%) | Spe.(%) | **Score**(%) |
|---|---|---|---|---|
| Naïve [18] | – | 48.34 | 64.28 | 56.31 |
| Noise | audio | 50.21 | 62.06 | 56.14 |
| Speed,loudness,shift | audio | 47.83 | 64.28 | 56.06 |
| Concat+Blank [15] | audio, spec | 54.46 | 78.53 | 66.50 |
| Mixup metric [19] | embed | 51.74 | **79.48** | 65.61 |
| Mixup [12] | spec | 55.88 | 71.82 | 63.85 |
| SpecAug w/o Mixup | spec | 50.89 | 77.96 | 64.43 |
| PatchMask w/o Mixup | spec | 54.88 | 76.18 | 65.53 |
| **GaP-aug w/o Mixup** | spec | **56.49** | 76.94 | **66.72** |
| SpecAug w/ Mixup | spec | 48.63 | **79.54** | 64.09 |
| PatchMask w/ Mixup | spec | 54.88 | 77.01 | 65.94 |
| **GaP-aug w/ Mixup** | spec | **58.20** | 77.07 | **67.64** |

### 3.3. Analysis

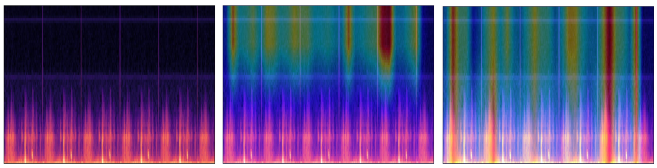
In this section, we present a visualization analysis to demonstrate the impact of our augmentation method on our backbone CNN model. We utilize method of Grad-CAM [20] to locate the critical parts within the testing spectrograms that play significant roles in the model's predictions. The process works by identifying and processing the gradient of the target sample in the final convolutional layer; Grad-CAM then provides a heat map, facilitating the visualization of how the model interprets our sample, and which specific areas within the images are most important on prediction outcomes. This technique assists in analyzing whether our predictions align with the medical descriptions of abnormal characteristics.

We conduct a comparative analysis of Grad-CAM results between our method and SpecAug [11], because SpecAug is a widely adopted augmentation that has a similar operation to our approach. Fig. 2a shows the outcome of an illustration sourced from the *Both* class. In the case of SpecAug, its emphasis is dominantly localized at the low-frequency and short-temporal segments of the spectrogram (the Grad-Cam visualization displays in red), while overlooking the high-frequency bands, results in wrong prediction of *Crackle*. Conversely, GaP-aug notices both low and high-frequency characteristics along with temporal continuity, resulting in correct prediction. Fig. 2b demonstrates an example sourced from the *Crackle* class. SpecAug is misdirected toward the high-frequency segments, not focusing on the critical low-frequency explosive sounds. By contrast, GaP-aug shows attention on information around the low-frequency intervals. Finally, Fig. 2c illustrates an instance sourced from the *Wheeze* class. In the context of SpecAug, it fails to detect abnormal features, thus leading to an incorrect prediction of *Normal*. On the contrary, GaP-aug demonstrates attention towards continuous segments within high-frequency bands.

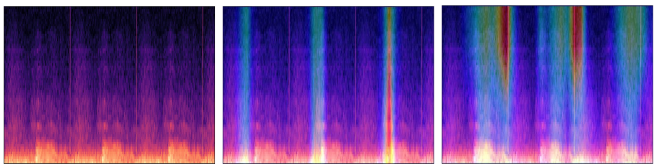
Our analysis points out an issue when implementing SpecAug on respiratory sound classification. Masking time and frequency intervals directly may cause misdirection during training process, since it is likely that critical acoustic signature of pathological breathing sounds may have been masked as well. Instead, our GaP-aug works by not only increasing the dataset diversity, but also preserves the distinct characteristics of abnormal sound classes.



(a) Grad-CAM example of *Both*, SpecAug predicted *Crackle*.



(b) Grad-CAM example of *Crackle*, SpecAug predicted *Wheeze*.



(c) Grad-CAM example of *Wheeze*, SpecAug predicted *Normal*.

**Fig. 2**: Grad-CAM of abnormal instances. The left column is original spectrograms, mid is overlay of Grad-CAMs from SpecAug, right is overlay of Grad-CAMs from GaP-aug.

## 4. CONCLUSION AND FUTURE WORK

In this work, we propose a gamma patch-wise correction augmentation technique, which involves contrast rescaling of random patches of spectrograms. GaP-aug effectively enhances performances in detecting abnormal classes by addressing two key challenges in respiratory sound classification: data scarcity and unique acoustic characteristic among abnormal classes. Our method achieves state-of-the-art results on the 60-40 official split and 80-20 split of 5-fold cross-validation on ICBHI dataset, outperforming all existing approaches. Furthermore, this technique is straightforward to implement while yielding significant improvements. Through Grad-CAM analysis, it shows that our approach helps guide the learning to focus on specific regions of the spectrogram aligns with the known medical descriptions of abnormal respiratory sound characteristics. In future, we intend to extend the application of our methodology to diverse datasets encompassing distinct types of medical acoustics, such as heart murmurs. Furthermore, we also like to extend our method to handle use cases in emergency department where real-world noise may further introduce unwanted distortion to the spectrograms.

## 5. REFERENCES


[1] T. Lancet, "GBD 2017: a fragile world," 2018.

[2] B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, *et al.*, "A respiratory sound database for the development of automated classification," in *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*, pp. 33–37, Springer, 2018.

[3] I. Moummad and N. Farrugia, "Pretraining respiratory sound representations using metadata and contrastive learning," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023.

[4] W. Song, J. Han, and H. Song, "Contrastive embeddind learning method for respiratory sound classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1275–1279, IEEE, 2021.

[5] D. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 50–55, IEEE, 2019.

[6] S. Bae, J.-W. Kim, W.-Y. Cho, H. Baek, S. Son, B. Lee, C. Ha, K. Tae, S. Kim, and S.-Y. Yun, "Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification," in *Proc. INTERSPEECH 2023*, pp. 5436–5440, 2023.

[7] Z. Zhao, Z. Gong, M. Niu, J. Ma, H. Wang, Z. Zhang, and Y. Li, "Automatic respiratory sound classification via multi-branch temporal convolutional network," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9102–9106, IEEE, 2022.

[8] T. Nguyen and F. Pernkopf, "Lung sound classification using co-tuning and stochastic normalization," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 9, pp. 2872–2882, 2022.

[9] 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, p. 17186, 2021.

[10] L. Xu, J. Cheng, J. Liu, H. Kuang, F. Wu, and J. Wang, "ARSC-Net: Adventitious respiratory sound classification network using parallel paths with channel-spatial attention," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1125–1130, IEEE, 2021.

[11] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. INTERSPEECH 2019*, pp. 2613–2617, 2019.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[13] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2778–2788, 2021.

[14] N. Shimano, "Suppression of noise effects in color correction by spectral sensitivities of image sensors," *Optical Review*, vol. 9, pp. 81–88, 2002.

[15] S. Gairola, F. Tom, N. Kwatra, and M. Jain, "RespireNet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 527–530, IEEE, 2021.

[16] Z. Wang and Z. Wang, "A domain transfer based data augmentation method for automated respiratory classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9017–9021, IEEE, 2022.

[17] D. Kumar *et al.*, "Multi spectral feature extraction to improve lung sound classification using cnn," in *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 186–191, IEEE, 2023.

[18] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[19] S. Venkataramanan, B. Psomas, E. Kijak, L. Amsaleg, K. Karantzalos, and Y. Avrithis, "It takes two to tango: Mixup for deep metric learning," in *ICLR 2022-10th International Conference on Learning Representations*, pp. 1–21, 2022.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.