# Adversarial Fine-tuning using Generated Respiratory Sound to Address Class Imbalance

**June-Woo Kim**[1,4], **Chihyeon Yoon**[1], **Miika Toikkanen**[2], **Sangmin Bae**[3,4], **Ho-Young Jung**[1]*
[1]Department of AI, Kyungpook National University
[2]ALI Co., Ltd, [3]KAIST AI, [4]RSC LAB, MODULABS
Republic of Korea
{kaen2891, chichi8969, hoyjung}@knu.ac.kr
miika.toikkanen.2@gmail.com, bsmn0223@kaist.ac.kr

## Abstract

Deep generative models have emerged as a promising approach in the medical image domain to address data scarcity. However, their use for sequential data like respiratory sounds is less explored. In this work, we propose a straightforward approach to augment imbalanced respiratory sound data using an audio diffusion model as a conditional neural vocoder. We also demonstrate a simple yet effective adversarial fine-tuning method to align features between the synthetic and real respiratory sound samples to improve respiratory sound classification performance. Our experimental results on the ICBHI dataset demonstrate that the proposed adversarial fine-tuning is effective, while only using the conventional augmentation method shows performance degradation. Moreover, our method outperforms the baseline by 2.24% on the ICBHI Score and improves the accuracy of the minority classes up to 26.58%. For the supplementary material, we provide the code at https://github.com/kaen2891/adversarial_fine-tuning_using_generated_respiratory_sound.

## 1 Introduction

Deep generative models (DGMs) have become popular due to their potential to address data scarcity via augmentation. Among recent advancements, methods such as generative adversarial networks (GANs) [11], variational autoencoders (VAEs) [18], and diffusion probabilistic models [14] are gaining attraction. Notably, previous studies [1, 4] have demonstrated that training a model on a mixture of real samples and synthetic samples generated by DGMs can be an effective approach to better utilize the limited data. In medical domain, DGMs have been successfully leveraged to synthesize medical data in a variety of categories, including retinal images [5, 15], CT and MRI scans [26, 30, 29], as well as X-rays [21, 22]. However, synthesis is more challenging when it comes to sequential medical data, including respiratory sound [19, 16, 28], due to its complex temporal dynamics, high dimensionality and relative lack of benchmarks.

In this paper, we aim to generate high-fidelity respiratory sound samples using DGMs and then combine these synthetic samples with real data to improve the respiratory sound classification task, especially for imbalanced lung sound disease classes. Figure 1 illustrates the overall process of our approach split into phases 1 and 2. In phase 1, we introduce a simple method for generating respiratory sound samples using a conditional neural vocoder inspired by the recent success of audio diffusion models [14, 20] in obtaining realistic audio. However, the discrepancy between synthetic and real samples can introduce problems related to distribution inconsistency, which can degrade the performance of respiratory sound classification models as the proportion of synthetic data in the

*corresponding author

1st Workshop on Deep Generative Models for Health at NeurIPS 2023.

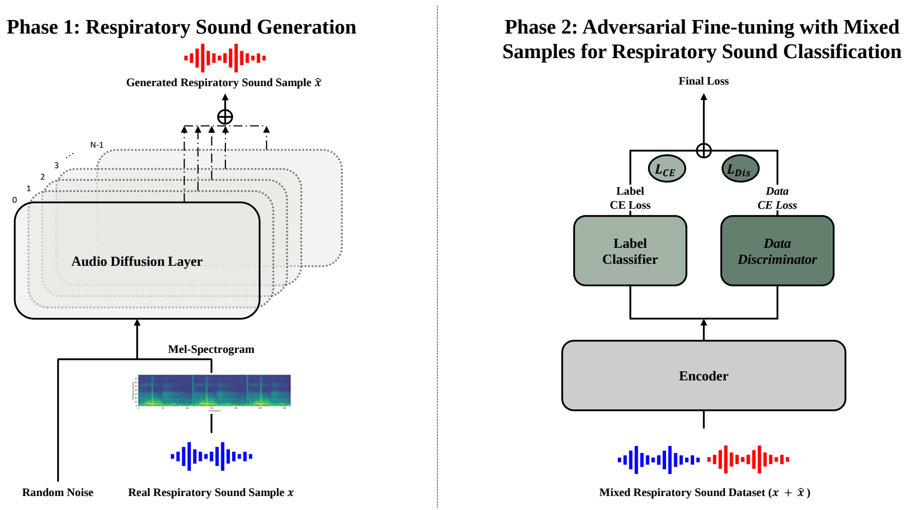



Figure 1: In phase 1, we generate the respiratory sound samples using the audio diffusion model as conditional neural vocoder. In phase 2, we use the proposed adversarial fine-tuning method to address the distribution inconsistency between synthetic and real samples for training the respiratory sound classification model.

training set increases. In phase 2, we propose a simple yet effective *adversarial fine-tuning* method motivated by [9] to learn the model that is distribution-agnostic between real and synthetic data. The adversarial fine-tuning method relies on a discriminator network feedback to move features obtained from real and synthetic samples closer to each other, while simultaneously training a classifier to predict the respiratory sound label.

Our experimental results on the ICBHI [27] dataset demonstrate that the proposed adversarial fine-tuning method effectively aligns the features from synthetic and real data, leading to improved performance while simply combining synthetic and real samples for training resulted in performance degradation. Specifically, our method achieves a 2.24% ICBHI Score improvement over the baseline and up to 26.58% accuracy improvement of the minority classes. Our contributions are: $(i)$ We show the successful generation of high-fidelity respiratory sound samples with audio diffusion model as conditional neural vocoder $(ii)$ We demonstrate adversarial fine-tuning on respiratory sound data, which can overcome data distribution inconsistency between synthetic and real samples $(iii)$ We present that the proposed method enables the synthetic and real training samples to be used more effectively, considerably improving performance in the imbalanced abnormal lung disease class.

## 2 Method

**Audio Diffusion Probabilistic Model** Diffusion probabilistic models [14] are a type of deep generative model that use a Markov chain to gradually add Gaussian noise $\mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ into a complex data distribution. The posterior $q(x_1, ..., x_T|x_0)$ called *diffusion process* or *forward process* is defined by a fixed Markov chain that transforms the input data $x_0$ to a latent variable $x_1, ..., x_T$ according to a variance schedule $\beta_1, ..., \beta_T$:

$$q_\theta(x_1, ..., x_T|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}), \qquad q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

The joint distribution $p_\theta(x_0, ..., x_{T-1}|x_T)$ called *reverse process* is defined by a Markov chain with learned Gaussian transitions starting at $p(x_T) = \mathcal{N}(x_T; 0, I)$:

$$p_\theta(x_0, ..., x_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{2}$$

where the transition probability $p_\theta(x_{t-1}|x_t)$ is parameterized as $\mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)^2 I)$ with shared parameter $\theta$, and the both $\mu_\theta$ and $\sigma_\theta$ are calculated with the diffusion-step and $x_t$.

In this work, we use the audio diffusion model [20] which is neural vocoding conditioned on Mel-spectrogram as a conditional neural vocoder to reconstruct the respiratory sound raw waveform. To this end, we employ the `DiffWave_BASE` model for our audio diffusion model, which consists of a stack of 30 residual layers, each with 64 residual channels as well as bidirectional dilated convolution with kernel size 3, and the dilation is doubled at each layer within each block. To ensure that the output of the audio diffusion model has the same length as the Mel-spectrogram, a transposed 2D convolution upsampler is provided for the conditioned 2D Mel-spectrogram.

In our conditional neural vocoder setting, the outputs of the upsampler are added to the dilated convolutions in each residual layer for reconstruction. In other words, our audio diffusion model is conditioned on the Mel-spectrogram, which means that it uses a lot of prior knowledge to guide the generation process. This makes it easier to generate realistic samples and reduces the need for large amounts of training data. This is especially beneficial in the medical domain, where data is often scarce.

**Adversarial Fine-tuning** While augmenting real data with synthetic samples can be beneficial, we found in early experiments that in our case the distribution mismatch between the two types of data degraded the performance of the classification model. To overcome this issue, we propose a simple yet effective **A**dversarial **F**ine-**T**uning (AFT) method inspired by [9]. The proposed method consists of two losses with label classifier $\mathcal{L}_{\text{CE}}$ and data discriminator $\mathcal{L}_{\text{Dis}}$:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{n} y_i \log(\hat{y}_i), \quad \mathcal{L}_{\text{Dis}} = -\frac{1}{N}\sum_{i=1}^{n} d_i \log(\hat{d}_i). \tag{3}$$

where $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{Dis}}$ are CE loss with label $y$ and data type label $d$, and the predicted probabilities $\hat{y}$ and $\hat{d}$ are obtained by passing through the classifier and data discriminator, respectively. To ensure that the learned features cannot distinguish between the synthetic and real samples, gradients from $\mathcal{L}_{\text{Dis}}$ are multiplied by a negative constant during the backpropagation. The final training objective is $\mathcal{L}_{\text{Final}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{Dis}}$ where $\lambda$ is a regularization parameter drawn from [9]. The AFT aims to reduce classification error while ensuring learned features are consistent across data types.

## 3 Experimental Setup

**ICBHI and Mixed Dataset** We used the ICBHI [27] dataset for respiratory sound tasks, following the official train-test split (60/40%). The training (4,142) and test sets (2,756) contain four classes: *normal* (49.8%/57.29%), *crackle* (29.3%/23.55%), *wheeze* (12.1%/13.97%) and *both* (8.8%/5.19%), hereinafter referred to as $C_n$, $C_c$, $C_w$ and $C_b$, respectively. We generated synthetic samples for each minority class to balance them with the majority class, and then mixed these with real data. We denote it *Mixed-ICBHI* datasets as follows: *Mixed-500, ..., Mixed-N, ..., Mixed-5k* where the number $N$ refers to the total amount of samples per class. We prioritize real samples so that synthetic samples are only added if the sample count is less than $N$. We used the *Specificity*, *Sensitivity* and their arithmetic mean, hereinafter referred to as $S_p$, $S_e$, and $Score$, respectively [27]. For ICBHI details and additional statistics on the Mixed-ICBHI dataset, see Appendix B and C.

**Audio Diffusion Model** For the data pre-processing, we fixed all of the data length as 4 seconds and extracted the 4,142 respiratory sound samples from the ICBHI dataset as 80-dimensional Mel-spectrograms. For the audio diffusion model, we trained the DiffWave [19] on the ICBHI training set from scratch. To this end, we used a linearly spaced schedule for the diffusion variance schedule parameter $\beta_t \in [1 \times 10^{-4}, 0.02]$, 50 and 6 denoising diffusion steps for training and evaluation, respectively. We then trained the model for 1M training steps with Adam [17] optimizer, a learning rate of 1e-4, and a batch size of 16.

**Respiratory Sound Classification** To prepare the data for training, we fixed the duration of all synthetic and real samples to 5 seconds and extracted 128-dimensional log Mel filterbank features with a window size of 25 ms and an overlap size of 10 ms. We then normalized the log Mel filterbank features using the mean and standard deviation of -4.27 and 4.57, as described in [3]. We trained the classification model using pretrained Audio Spectrogram Transformer [10] (AST) model with the Adam optimizer, a learning rate of 5e-5, and a batch size of 32 for 50 epochs. To ensure the stability of our results, we trained our model using a fixed set of five random seeds for all experiments.

Table 1: Respiratory sound classification performance on ICBHI test set according to various mixed sample amounts using the AST [10] fine-tuning as described in [3]. No Aug. denotes only the real ICBHI dataset is used for training. We only report the ICBHI *Score* (%). **Bold** denotes the best result.

| method | training dataset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | No Aug. | Mixed-500 | Mixed-800 | Mixed-1k | Mixed-1.5k | Mixed-2k | Mixed-3k | Mixed-5k |
| AST FT | 59.55 | $59.92_{\pm 0.82}$ | $59.99_{\pm 1.17}$ | $59.81_{\pm 0.36}$ | $59.65_{\pm 0.30}$ | $59.18_{\pm 0.65}$ | $59.04_{\pm 0.32}$ | $58.56_{\pm 0.84}$ |
| AFT | - | $\mathbf{61.79}_{\pm 0.47}$ | $60.89_{\pm 0.78}$ | $60.8_{\pm 1.05}$ | $60.03_{\pm 1.14}$ | $60.64_{\pm 0.45}$ | $59.96_{\pm 0.38}$ | $59.74_{\pm 0.6}$ |

Table 2: Accuracy (%) of the abnormal class on the ICBHI test set for AST fine-tuning and AST adversarial fine-tuning models trained on different datasets. **Bold** denotes the best result.

| class | ratio | method (dataset) | | | | |
|---|---|---|---|---|---|---|
| | | AST FT (No aug.) | AST FT (Mixed-500) | AST FT (Mixed-2k) | Adversarial FT (Mixed-500) | Adversarial FT (Mixed-2k) |
| crackle ($C_c$) | 23.55% | 45.45 | 42.84 | 42.86 | 44.07 | **46.99** |
| wheeze ($C_w$) | 13.97% | 36.62 | 36.1 | 22.08 | **37.92** | 30.12 |
| both ($C_b$) | 5.19% | 15.38 | 9.09 | 7.69 | **41.96** | 35.66 |

Table 3: Overall comparison of the ICBHI dataset for the respiratory sound classification task. We compared previous studies that followed the official 60-40% split for the training/test set. Scores marked with $*$ denote the previous state-of-the-art performance. **Best** and <u>second best</u> results.

| method | architecture | pretrain | $S_p$ (%) | $S_e$ (%) | **Score** (%) |
|---|---|---|---|---|---|
| RespireNet [8] (CBA+BRC+FT) | ResNet34 | IN | 72.30 | 40.10 | 56.20 |
| Wang *et al.* [32] (Splice) | ResNeSt | IN | 70.40 | 40.20 | 55.30 |
| Nguyen *et al.* [25] (CoTuning) | ResNet50 | IN | 79.34 | 37.24 | 58.29 |
| Moummad *et al.* [23] (SCL) | CNN6 | AS | 75.95 | 39.15 | 57.55 |
| Bae *et al.* [3] (Fine-tuning) | AST | IN + AS | 77.14 | 41.97 | 59.55 |
| Bae *et al.* [3] (Patch-Mix CL) | AST | IN + AS | **81.66** | **43.07** | **62.37***  |
| **AFT on Mixed-500 [ours]** | AST | IN + AS | $\underline{80.72}_{\pm 0.99}$ | $\underline{42.86}_{\pm 1.3}$ | $\underline{61.79}_{\pm 0.47}$ |

# 4 Results

## 4.1 Effectiveness of Adversarial Fine-tuning

To validate the proposed AFT, we compared it against AST fine-tuning (AST FT) with only cross-entropy (CE) loss on several Mixed-ICBHI datasets under the same conditions. As in Table 1, the AST FT performance decreased as the number of augmented samples in the ICBHI dataset increased, while the AFT outperformed it in each case, reaching the best Score on *Mixed-500*. Based on the result, as $N$ increases, the distribution mismatch between synthetic and real samples increases, therefore leading to reduced performance. Our method mitigates this to a degree, but still benefits more in smaller $N$. We further explore how our method affects the performance of minority classes. We report their accuracy on the ICBHI test set for AST FT with no augmentation, and AST FT and AFT on Mixed-500 and Mixed-2k. As in Table 2, directly fine-tuning on mixed data did not improve the performance of the minority classes overall. However, our proposed method improved their accuracies by up to 26.58%, especially in $C_b$. These results show that our method can most effectively enhance the performance of minority classes despite using synthetic samples that would otherwise degrade them. For additional confusion matrices of Table 2, see Appendix D.

## 4.2 Comparison on ICBHI Dataset Results

Table 3 presents an overall comparison of various methods for lung sound classification on the ICBHI dataset. Our proposed method trained with Mixed-500 achieved a Score of 61.79%, outperforming the AST FT model by 2.24%, which is comparable to the state-of-the-art model. This demonstrates the efficacy and potential of our proposed method, indicating its capability for addressing the issues with synthetic data.

### 4.3 Qualitative Analysis

Figure 2 provides visual comparison of spectrograms randomly sampled per class from the test set and the results generated by our diffusion model when conditioned on these spectrograms. The generated spectrograms per class are visually similar to the original sample, which demonstrates the capability to generate high-fidelity audio, yet introduce small realistic variations that provide some value for augmentation.

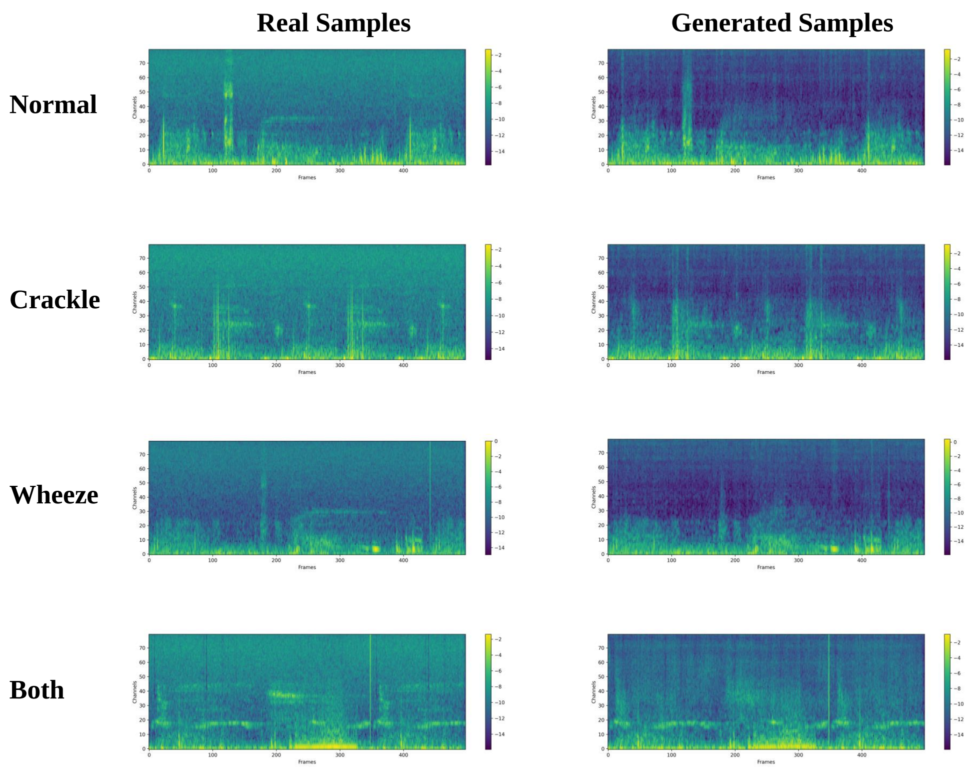



Figure 2: Comparison of spectrograms per each class randomly chosen from the test set and the generated results.

## 5 Conclusion

We presented a simple method for generating realistic respiratory sound samples using an audio diffusion model. We further introduced adversarial fine-tuning to address the distribution inconsistency between synthetic and real samples. Our results show that our method can effectively improve the performance of imbalanced abnormal classes, demonstrating its ability to address the challenges of using synthetic data. We believe that our method can be helpful in various other datasets and could be used to supplement other augmentation methods.

## Acknowledgments and Disclosure of Funding

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2023-2020-0-01808) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation), and by Brian Impact, a non-profit organization dedicated to the advancement of science and technology.

## References


[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

[2] Karim Armanious, Chenming Jiang, Marc Fischer, Thomas Küstner, Tobias Hepp, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, 79:101684, 2020.

[3] Sangmin Bae, June-Woo Kim, Won-Yang Cho, Hyerim Baek, Soyoun Son, Byungjo Lee, Changwan Ha, Kyongpil Tae, Sungnyun Kim, and Se-Young Yun. Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification. In *Proc. INTERSPEECH 2023*, pages 5436–5440, 2023.

[4] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

[5] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017.

[6] Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*, 2022.

[7] Fatemeh Fahimi, Zhuo Zhang, Wooi Boon Goh, Kai Keng Ang, and Cuntai Guan. Towards eeg generation using gans for bci applications. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE, 2019.

[8] Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit 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)*, pages 527–530. IEEE, 2021.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[10] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[12] Changhee Han, Yoshiro Kitamura, Akira Kudo, Akimichi Ichinose, Leonardo Rundo, Yujiro Furukawa, Kazuki Umemoto, Yuanzhong Li, and Hideki Nakayama. Synthesizing diverse lung nodules wherever massively: 3d multi-conditional gan-based ct image augmentation for object detection. In *2019 International Conference on 3D Vision (3DV)*, pages 729–737. IEEE, 2019.

[13] Debapriya Hazra and Yung-Cheol Byun. Synsiggan: Generative adversarial networks for synthetic biomedical signal generation. *Biology*, 9(12):441, 2020.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[15] Talha Iqbal and Hazrat Ali. Generative adversarial network for medical images (mi-gan). *Journal of medical systems*, 42:1–11, 2018.

[16] S Jayalakshmy and Gnanou Florence Sudha. Conditional gan based augmentation for predictive modeling of respiratory signals. *Computers in Biology and Medicine*, 138:104930, 2021.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Kirill Kochetov and Andrey Filchenkov. Generative adversarial networks for respiratory sound augmentation. In *Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System*, pages 106–111, 2020.

[20] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

[21] Mohamed Loey, Florentin Smarandache, and Nour Eldeen M. Khalifa. Within the lack of chest covid-19 x-ray dataset: a novel detection model based on gan and deep transfer learning. *Symmetry*, 12(4):651, 2020.

[22] Saman Motamed, Patrik Rogalla, and Farzad Khalvati. Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images. *Informatics in Medicine Unlocked*, 27:100779, 2021.

[23] Ilyass Moummad and Nicolas Farrugia. Supervised contrastive learning for respiratory sound classification. *arXiv preprint arXiv:2210.16192*, 2022.

[24] Pedro Narváez and Winston S Percybrooks. Synthesis of normal heart sounds using generative adversarial networks and empirical wavelet transform. *Applied Sciences*, 10(19):7003, 2020.

[25] Truc Nguyen and Franz Pernkopf. Lung sound classification using co-tuning and stochastic normalization. *IEEE Transactions on Biomedical Engineering*, 69(9):2872–2882, 2022.

[26] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 417–425. Springer, 2017.

[27] BM Rocha, Dimitris Filos, L Mendes, Ioannis Vogiatzis, Eleni Perantoni, E Kaimakamis, P Natsiavas, Ana 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*, pages 33–37. Springer, 2018.

[28] Jane Saldanha, Shaunak Chakraborty, Shruti Patil, Ketan Kotecha, Satish Kumar, and Anand Nayyar. Data augmentation using variational autoencoders for improvement of respiratory disease classification. *Plos one*, 17(8):e0266467, 2022.

[29] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):16884, 2019.

[30] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 1–11. Springer, 2018.

[31] Chenxi Tian, Yuliang Ma, Jared Cammon, Feng Fang, Yingchun Zhang, and Ming Meng. Dual-encoder vae-gan with spatiotemporal features for emotional eeg data augmentation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[32] Zijie Wang and Zhao 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)*, pages 9017–9021. IEEE, 2022.

[33] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022.

[34] Naren Wulan, Wei Wang, Pengzhong Sun, Kuanquan Wang, Yong Xia, and Henggui Zhang. Generating electrocardiogram signals by deep learning. *Neurocomputing*, 404:122–136, 2020.

## A Related Works

**Respiratory Sound Classification** The ICBHI [27] dataset is a well-known benchmark for respiratory sound classification. Various neural network-based approaches have been developed for this task, including residual blocks [8, 25, 32], CNN [23], pretrained models on ImageNet [8, 25, 32], AudioSet [23], and Audio Spectrogram Transformer (AST) [10]. To address the challenge of limited data, previous studies have proposed various learning protocols, including device-specific fine-tuning [8], mixup as well as splicing audio augmentation [32], task-specific co-tuning [25], supervised contrastive learning [23], and patch-mix contrastive learning [3]. Instead of focusing on previous data augmentation methods, this paper addressed the challenge of using synthetic samples generated by deep generative models. To this end, we first trained a pre-trained AST [3] model on the ICBHI dataset as described in [10]. We also trained the model on the Mixed-ICBHI dataset, which contains both synthetic and real samples. We then showed that our proposed adversarial fine-tuning method can overcome the data distribution inconsistency between synthetic and real samples.

**Deep Generative Models** Recent advances in DGMs, such as GAN [11], VAE [18], and diffusion models [14], have attracted significant attention. This is because DGMs can be used to generate synthetic samples to mitigate data scarcity issues. They have been applied to medical images, such as retinal images [5, 15], CT and MRI scans [26, 30, 4, 29, 12, 2, 33, 6], and X-rays [21, 22] which received additional interest due their applicability in diagnosing COVID-19 cases. Several approaches have also been introduced to generate synthetic sequential medical data, such as respiratory sounds [19, 16, 28], EEG recordings [7, 13, 31], and ECG signals [24, 34]. Unlike previous studies on respiratory sound, our work was the first attempt to successfully generate high-fidelity respiratory sound samples using an audio diffusion model [20] which is neural vocoding conditioned on Mel-spectrogram as a conditional neural vocoder.

## B ICBHI Dataset Details

Table 4: Overall details of the ICBHI [27] respiratory sound dataset.

| | number of respiratory samples (ratio) | | |
|---|---|---|---|
| label | train | test | sum |
| Normal | 2,063 (49.8%) | 1,579 (57.29%) | 3,642 |
| Crackle | 1,215 (29.3%) | 649 (23.55%) | 1,864 |
| Wheeze | 501 (12.1%) | 385 (13.97%) | 886 |
| Both | 363 (8.8%) | 143 (5.19%) | 506 |
| Total | 4,142 | 2,756 | 6,898 |

Table 5: Overall details of Mixed-ICBHI dataset with synthetic and real samples.

| | mixed dataset (synthetic ratio, %) | | | | | | |
|---|---|---|---|---|---|---|---|
| label | Mixed-500 | Mixed-800 | Mixed-1k | Mixed-1.5k | Mixed-2k | Mixed-3k | Mixed-5k |
| normal | 0 | 0 | 0 0 | 0 | 0 | 31.23 | 58.74 |
| crackle | 0 | 0 | 0 | 19.00 | 41.11 | 59.50 | 75.70 |
| wheeze | 0 | 37.38 | 49.90 | 66.60 | 75.72 | 83.30 | 89.98 |
| both | 27.40 | 54.63 | 63.70 | 75.80 | 82.40 | 87.90 | 92.74 |

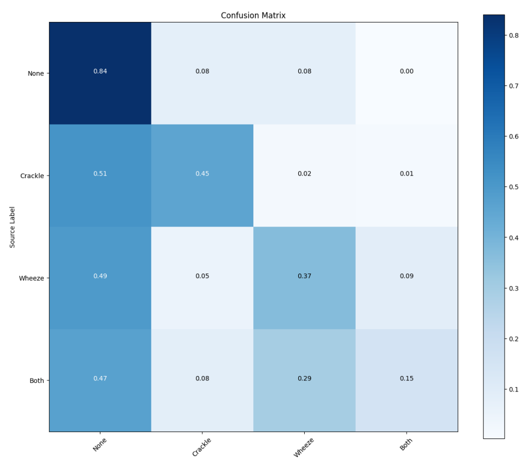
The ICBHI [27] dataset is a well-known benchmark for respiratory sound classification. The ICBHI dataset consists of 6,898 respiratory cycles, with a total duration of approximately 5.5 hours. The dataset is officially split into a training set (60%) and a test set (40%), with no patient overlap between the two sets. As shown in Table 4, the training and test sets contain 4,142 and 2,756 samples respectively and are categorized into four classes, *normal* (49.8%/57.29%), *crackle* (29.3%/23.55%), *wheeze* (12.1%/13.97%) and *both* (8.8%/5.19%), respectively. For all our experiments, we resampled all the samples to 16 kHz. For the metrics, we used *Sensitivity* ($S_e$), *Specificity* ($S_p$), and their arithmetic mean *Score* as described in [27].

## C Mixed Dataset Details

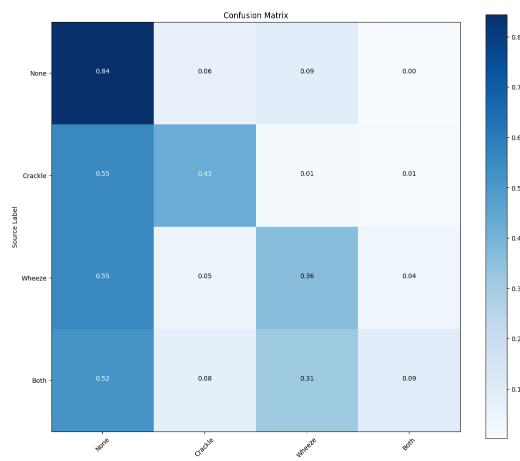
As described in Table 5, we mixed the synthetic samples with the real data to create Mixed-ICBHI datasets as follows: *Mixed-500, ..., Mixed-N, ..., Mixed-2k* where the number $N$ refers to the total amount of samples per class. We prioritize real samples so that synthetic samples are only added if the sample count is less than $N$ (i.e., *Mixed-500* only contains synthetic samples from $C_{both}$).

## D Confusion Matrices Results

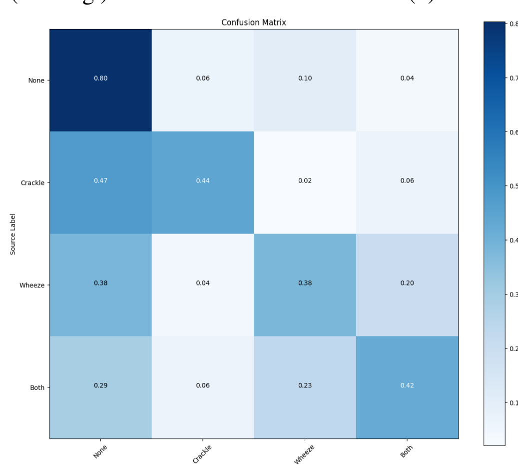
To show how the proposed method affects *all the classes*, Figure 3 provides the confusion matrices between the AST FT with no augmentation, AST FT and AFT with Mixed-500 and Mixed-2k, respectively. The proposed method did not degrade considerably on normal classes and achieved the highest performance compared to other methods on the most imbalanced classes. Our results demonstrate the effectiveness and potential of our proposed method, showing its ability to address the data distribution inconsistency problem with synthetic data, especially in class imbalanced problems.




(a) AST FT (No Aug.)




(b) AST FT (Mixed-500)




(c) AFT (Mixed-500)

Figure 3: Confusion matrix results of AST FT with no augmentation, AST FT and AFT with Mixed-500 and Mixed-2k, respectively.