PASA: Progressive-Adaptive Spectral Augmentation for Automated Auscultation in Data-Scarce Environments

Ying Wang[1], Guoheng Huang[2], Xueyuan Gong[3], Xinxin Wang[4], Xiaochen Yuan[1,*]
[1]Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, China
[2]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China
[3]School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai, China
[4]School of Artificial Intelligence, Shenzhen University, Shenzhen, China
p2412918@mpu.edu.mo, kevinwong@gdut.edu.cn, xygong@jnu.edu.cn, xinxinwang1024@gmail.com, xcyuan@mpu.edu.mo

Abstract

Automated auscultation advances the detection of respiratory diseases, especially in areas with limited resources where traditional diagnostic methods are unavailable. On the other hand, the scarcity of auscultation datasets limits the automation performance, prompting the needs for data augmentation methods. However, most of the existing methods neglect the difference in acoustic sounds that requires personalized augmentation strategies. To address this, we propose a Progressive-Adaptive Spectral Augmentation (PASA), which is one of the first paradigms to adaptively select the best augmentation strategy for each sample. The PASA innovatively treats augmentation selection problem as a Markov Decision Process (MDP), creating an alternating loop between the diagnostic model and the augmentation selection. The agent selects the optimal augmentation operations and magnitudes via a task-specific design, including state construction, action sampling, Hybrid Batch-Sample (HBS) strategy execution, and reward guidance. The HBS strategy initially applies uniform augmentation across mini-batches while collecting sample-specific performance statistics. When model performance stabilizes, it transits to sample-level augmentation based on accumulated difficulty assessments. This two-phase design balances computational complexity with personalization. Extensive experiments across three benchmark datasets demonstrate that the PASA outperforms the state-of-the-art methods, pioneering a transformative paradigm for adaptive data augmentation in automated auscultation.

Code — https://github.com/wangying1586/PASA

Introduction

Conditions such as asthma and cardiovascular disease produce abnormal respiratory sounds, posing a threat to global health and highlighting the importance of early detection (Al-Hasan et al. 2025). However, current detection methods rely on physician expertise and expensive imaging, which are often unavailable due to financial and infrastructural limitations (Bernardi et al. 2019). To address this, researchers began exploring automated, computer-aided auscultation (Osama et al. 2024). These automated methods show promise for telemedicine, routine screening programs, and use in underserved areas. Automated auscultation has the potential to transform the delivery of respiratory healthcare across various clinical settings (Heitmann et al. 2023). The success of such automation relies on effective audio preprocessing. Preprocessed spectral features are ideal for detecting the frequency patterns that show various respiratory conditions (Park et al. 2025).

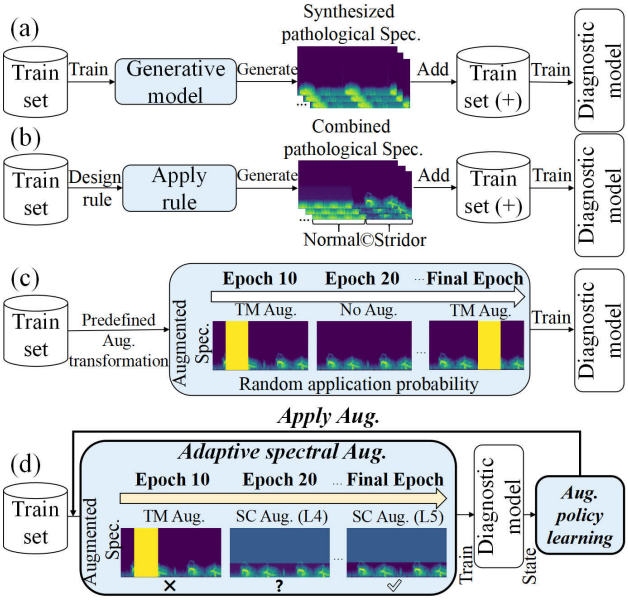

Figure 1: Illustration of the differences among (a) generative methods, (b) combination rule-based methods, (c) transformation predefined methods, and (d) our PASA.

However, training effective spectral-based models requires large amounts of labeled datasets (Whang et al. 2023). Unfortunately, such datasets remain scarce in the auscultation domain (Zhao et al. 2024). This is because medical data collection is expensive, and medical experts are needed to validate the data, especially for rare respiratory conditions (Ren et al. 2024). To address this, researchers have turned to developing data augmentation methods, as illustrated in Figure 1. Specifically, in Figure 1 (a), some employ generative models (Kim et al. 2023) to synthesize artificial samples. In Figure 1 (b), others utilize prior knowl-

*Corresponding author

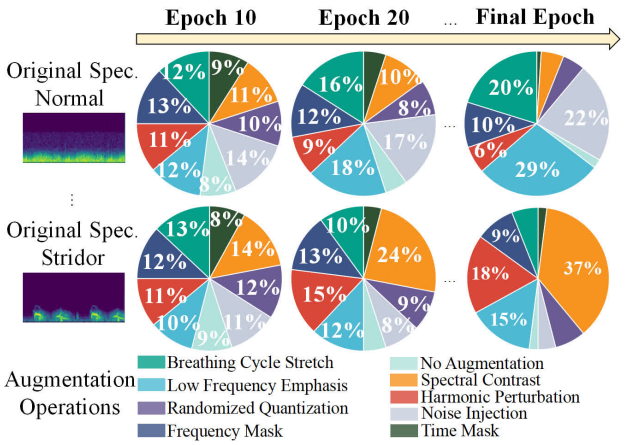

Figure 2: Augmentation operation evolution in our proposed PASA showing transition from uniform distributions to sample-specific strategies.

edge to design rules for combining sound segments (Wang and Wang 2022; Wang et al. 2024a,b). Both methods generate the pathological spectrograms to be added to the training data. In most cases, in Figure 1 (c), methods are based on applying fixed transformations with set probabilities during training (Liu et al. 2022; Pessoa et al. 2023; Ngo et al. 2023a; Hu et al. 2025). For example, during the training period, the time masking (TM Aug.) (Park et al. 2019) pre-sets a time amplitude of 0.2 to mask the spectral information. This process is done randomly, with a 0.5 probability of enabling or disabling. These static methods are applied equally to all samples across training epochs, regardless of the characteristics of each respiratory spectrogram. This makes them unable to find the most suitable augmentation strategy for each sample and fully utilize the scarce training data.

Due to the limitations of static augmentation, recent studies have started exploring reinforcement learning (RL)-based adaptive augmentation. These RL-driven methods are commonly used in computer vision to improve image data, but less so for audio analysis (Sun et al. 2024). These methods operate at the batch or sample levels. However, both methods face limitations when applied to auscultation tasks. First, batch-level methods (Müller and Hutter 2021; Hou, Zhang, and Zhou 2023) apply the same augmentations to each mini-batch to achieve computational efficiency. However, homogeneity issues arise when the same augmentations are applied to diverse samples, yielding suboptimal results. Second, sample-level methods (Yu et al. 2024; Li et al. 2025) apply different augmentations for each sample to improve personalization. However, they face several challenges. Cold-start problems occur due to inadequate initial data, reward fluctuations cause unstable training, and per-sample optimization incurs high computational costs. These issues reflect an efficiency-personalization trade-off that prevents the application of these methods to automatic auscultation tasks. A hybrid method that balances both factors is needed.

To address these issues, we propose a Progressive-Adaptive Spectral Augmentation (PASA), which dynamically learns optimal augmentation policies via RL feedback. As shown in Figure 1 (d), unlike static methods, PASA creates an adaptive learning process where unsuitable augmentations are gradually filtered out while effective ones are encouraged. For example, when an agent applies unsuitable augmentation like TM Aug. (Park et al. 2019) to stridor samples at epoch 10, negative feedback guides the agent to avoid such operations in future decisions like epoch 20. This iterative learning gradually filters out harmful augmentations. At the final epoch, it selects effective ones like spectral contrast (SC Aug.) with appropriate magnitude levels (L5). To better show the adaptive evolution in different samples, as shown in Figure 2. At epoch 10, an agent explores by trying different augmentations and measuring the impact on diagnostic performance. The results show that both samples have a similar usage rate of 10-14%. Through training and decision optimization, normal samples learn to prefer Low Frequency Emphasis, Noise Injection, and Breathing Cycle Stretch operations. These strategies make breathing more robust while keeping it natural. On the other hand, stridor samples show Spectral Contrast (37% usage) and Harmonic Perturbation (18% usage) at the final epoch. These operations make pathological acoustic features louder.

To realize this adaptive learning capability, PASA implements a five-step workflow including state construction, action sampling, Hybrid Batch-Sample (HBS) strategy execution, reward computation, and joint optimization. The core innovation lies in our Hybrid Batch-Sample (HBS) strategy, which serves as the execution component to intelligently balance computational efficiency and personalization via a two-phase design. Our main contributions are:

- We present one of the first adaptive augmentation paradigm based on RL for automatic auscultation, namely Progressive-Adaptive Spectral Augmentation (PASA). It automatically selects the optimal augmentations for each sample using a five-step workflow design.
- We introduce a Hybrid Batch-Sample (HBS) strategy as the augmentation executor of PASA, which starts with batch-level uniform augmentation to collect sample statistics, then transitions to sample-level personalized augmentation based on difficulty assessment, effectively balancing efficiency and personalization via a two-phase design.
- We achieve SOTA performance with improvements, including a 2% W.acc increase on the CirCor DigiScope and up to 12.6% improvement over recent SOTA on SPRSound 2023 tasks. This shows the success of adaptive augmentation in situations where data are limited, and its potential application to other medical AI areas.

Related Work

Related work includes data augmentation methods developed for automated auscultation and adaptive augmentation methods from computer vision domains.

Data augmentation methods

Data augmentation methods address the scarcity of data in automatic auscultation (Osama et al. 2024). Early generative models (Kim et al. 2023) create artificial spectrograms, but they generate pathological features in an uncontrollable manner. To address this limitation, rule-based combination methods (Wang and Wang 2022; Wang et al. 2024a,b) arose to provide better control by mixing audio segments, yet they remain constrained by the availability of original datasets. Therefore, most of methods (Liu et al. 2022; Pessoa et al. 2023; Ngo et al. 2023a; Hu et al. 2025) use spectral transformations for online data synthesis with fixed augmentation transformations. However, these static transformations fail to capture the distinct acoustic characteristics of normal versus abnormal respiratory sounds. Unlike these methods, PASA uses a RL process to select specific augmentations for each sample based on its unique acoustic properties.

Adaptive augmentation methods

Adaptive augmentation methods have been developed to overcome the limitations of static augmentation (Fawzi et al. 2016). Early work (Müller and Hutter 2021) used random selection from predefined ranges, which sacrifices true adaptability for computational simplicity. To achieve true adaptability, two main categories of methods have emerged. Batch-level methods (Hou, Zhang, and Zhou 2023) apply uniform strategies across mini-batches for computational efficiency. They assume samples within batches have similar augmentation needs. This assumption fails for acoustic data where normal and pathological sounds have different spectral characteristics. Sample-level methods (Yu et al. 2024; Li et al. 2025) provide personalized augmentation based on individual sample characteristics. While theoretically sound, per-sample optimization creates unmanageable complexity, leading to unstable training and high costs. Current methods are too rigid to allow dynamic granularity selection. They either favor personalization over efficiency or efficiency over personalization. Therefore, our HBS strategy dynamically adjusts adaptation granularity based on training progress and sample characteristics. It switches between batch-level and sample-level strategies through feedback, balancing efficiency with personalization.

Methodology

Preliminary

Adaptive augmentation methods employ RL to formulate augmentation selection as a MDP (Sutton and Barto 1998). The MDP consists of states S, actions A, transition probabilities P, rewards R, and discount factor $\gamma \in (0, 1]$. During each training step t, the agent receives state s_t, applies policy $\pi(a_t|s_t)$ to select action a_t, and obtains reward r_t and next state s_{t+1}. The objective is to optimize policy π^* that maximizes cumulative reward $R_t = \sum_{k=0}^{T} \gamma^k r_{t+k}$, where k denotes the time step index, T is the total number of time steps, and r_{t+k} represents the reward at future step $t + k$. In this work, we implement this MDP using Soft Actor-Critic (SAC) (Haarnoja et al. 2018) with actor network π_ϕ for policy learning and twin critics Q_{ψ_1}, Q_{ψ_2} for value estimation.

Algorithm 1: The main workflow of PASA

Input: Training mini-batch $\mathcal{B} = \{(x_j, y_j)\}_{j=1}^{B}$, validation set $\mathcal{D}_{val}$
Parameter: Diagnostic model f_θ, SAC networks including actor network π_ϕ and twin critic networks Q_{ψ_1}, Q_{ψ_2}
Output: Updated diagnostic model parameters and SAC networks

for each training epoch **do**
  for each mini-batch $\mathcal{B}$ **do**
    Extract features and construct state $\mathbf{s}_t$
    Sample augmentation action $\mathbf{a}_t \sim \pi_\phi(\mathbf{s}_t)$
    Execute HBS strategy to augment batch $\tilde{\mathcal{B}}$
    Evaluate on validation set $\mathcal{D}_{val}$ and compute r_t
    Update state $\mathbf{s}_{t+1}$ and jointly optimize networks
  end for
end for
return Updated parameters $\theta, \{\phi, \psi_1, \psi_2\}$

Overview

Based on the MDP formulation, we present our Progressive-Adaptive Spectral Augmentation (PASA) paradigm. As shown in Figure 3, PASA first preprocesses respiratory audio to generate log-mel spectrograms, then performs a split of training and validation sets, and adopts balanced mini-batch sampling. PASA implements a five-step workflow where each mini-batch serves as a discrete MDP time step, achieving dynamic adaptation during training.

Under this MDP framework, each training iteration follows a standard RL loop: $s_t \rightarrow a_t \rightarrow r_t \rightarrow s_{t+1}$. As shown in Algorithm 1, this loop is implemented via five steps. Firstly, **state construction** extracts semantic features and class distributions to form the MDP state s_t. Secondly, **action sampling** uses the SAC network to select augmentation policies as the MDP action a_t. Thirdly, **Hybrid Batch-Sample (HBS) execution strategy** performs selected augmentations with dynamic phase transitions. Fourthly, **reward calculation** evaluates performance improvement to generate the reward signal r_t. Finally, **joint optimization** updates both diagnostic model and SAC networks using experience tuples (s_t, a_t, r_t, s_{t+1}).

Progressive-Adaptive Spectral Augmentation

PASA begins with data preprocessing. Raw respiratory audios undergo resampling, denoising, normalization, and length standardization. The processed audios are then converted into log-mel spectrogram representations. The spectrogram dataset then is divided into training and validation sets using an 8:2 ratio. During training, PASA operates on mini-batches of log-mel spectrograms $\mathcal{B} = \{(x_j, y_j)\}_{j=1}^{B}$. Here, $x_j \in \mathbb{R}^{1\times H\times W}$ represents spectrograms and $y_j \in \{0, 1, ..., C-1\}$ denotes class labels. C is the number of classes. A balanced sampler processes the training set to generate class-balanced mini-batches. These mini-batches are used for the subsequent five-step workflow.

Feature extraction and state construction. For each mini-batch $\mathcal{B} = \{(x_j)\}_{j=1}^{B}$ at time step t, PASA first ex-

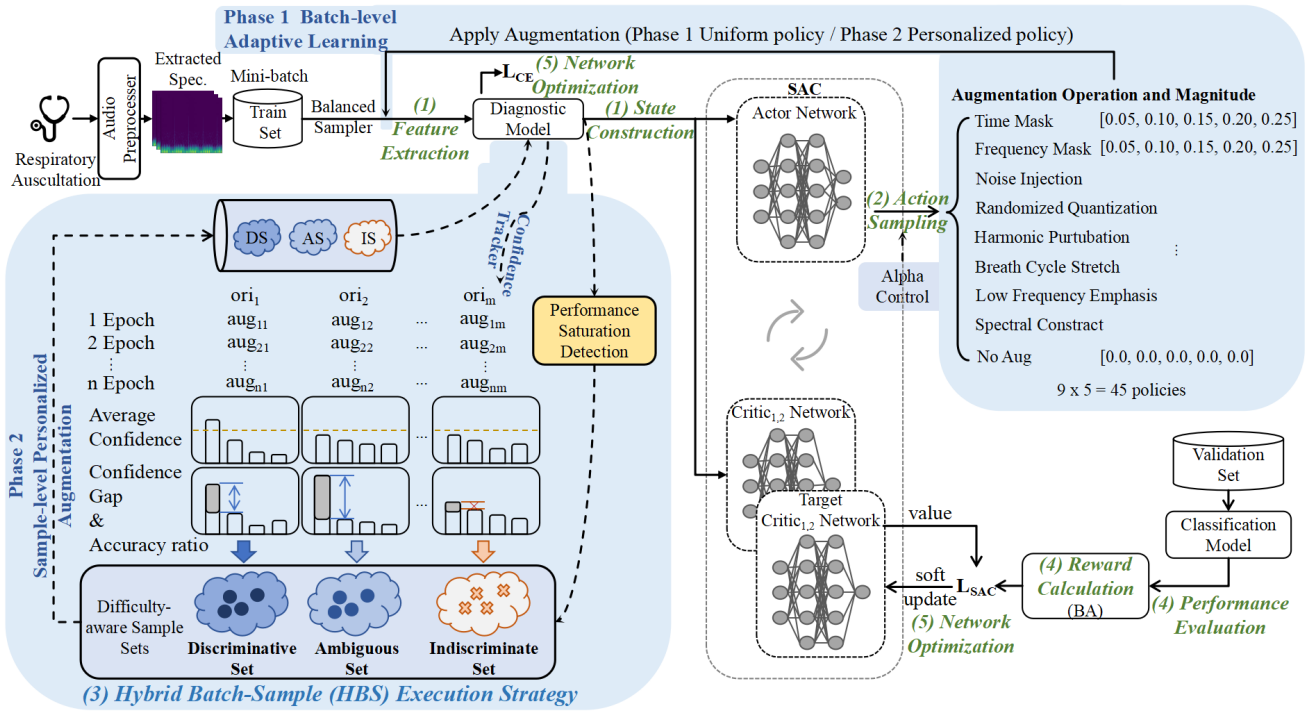

Figure 3: PASA paradigm with five-step workflow begin with the preprocessed datasets. Then, (1) feature extraction and state construction, (2) action sampling via SAC network, (3) HBS execution strategy, (4) performance evaluation and reward computation, and (5) joint optimization of augmentation policies and diagnostic models.

tracts semantic features $\mathbf{f}_{batch}$ from the diagnostic model f_θ via a global average pooling GAP:

$$\mathbf{f}_{batch} = \frac{1}{B}\sum_{j=1}^{B} \text{GAP}(f_\theta^{(L-1)}(x_j)) \in \mathbb{R}^d \quad (1)$$

where L is the total number of network layers, and d is the extracted feature dimension.

The extracted features $\mathbf{f}_{batch}$ are then combined with label category distribution $\mathbf{d}_{class}$ to construct the MDP state $\mathbf{s}_t$:

$$\mathbf{s}_t = [\mathbf{f}_{orig}, \mathbf{f}_{aug}, \mathbf{d}_{class}] \in \mathbb{R}^{2d+C} \quad (2)$$

where $\mathbf{f}_{orig}$ and $\mathbf{f}_{aug}$ are the semantic features extracted from the diagnostic model for the original and augmented versions of the current batch, respectively. At the first step $t = 0$, $f_{aug} = f_{orig}$ since no augmentation has been applied. For later steps, $\mathbf{f}_{aug}$ is the features of the current batch after the selected augmentation strategy has been applied.

Action sampling via SAC network. After state construction, PASA uses nine spectral operations (Abayomi-Alli et al. 2022) (Time Mask, Frequency Mask, Noise Injection, Harmonic Perturbation, Breath Cycle Stretch, Low Frequency Emphasis, Spectral Contrast, No Augmentation, Randomized Quantization (Wu et al. 2023)) with five magnitude levels, creating 45 policies. In SAC, the actor network π_ϕ takes the state $\mathbf{s}_t$ as input and outputs probabilities of augmentation operations $p_{op} \in \mathbb{R}^9$ and magnitudes $p_{mag} \in \mathbb{R}^5$. Categorical sampling then selects actions (op_{idx}, mag_{idx}) from these probability distributions.

Hybrid batch-sample execution strategy. After action sampling, HBS executes the selected actions (op_{idx}, mag_{idx}) on mini-batch $\mathcal{B}$ to output the augmented batch $\tilde{\mathcal{B}}$. As shown in Figure 3, HBS corresponds to the blue-highlighted part that bridges action sampling and performance evaluation in the PASA workflow.

HBS uses a two-phase design, where the phase switching is controlled by monitoring the performance at the epoch level. As shown in Algorithm 2, in Phase 1, HBS uses uniform augmentation across batches and updates sample statistics. In Phase 2, HBS applies Gaussian noise to prevent overfitting. Then it classifies sample difficulty for each sample in each batch and then samples personalized augmentation actions. Finally, it uses sample-level augmentation to each sample.

HBS monitors training progress to decide when to switch phases. After each epoch, HBS evaluates the performance of diagnostic model on the validation set. Phase transition occurs when validation performance shows no improvement for P consecutive epochs:

$$\text{Phase 2} \leftarrow \sum_{k=1}^{P} \mathbf{1}[BA_{val}^{(t-k)} \leq BA_{hist}^{best} + \delta] \geq P \quad (3)$$

where t is the current epoch, P denotes patience epochs, $BA_{val}^{(t-k)}$ represents the validation balanced accuracy at epoch $(t-k)$, i.e., k epochs before the current epoch, BA_{hist}^{best} is historical best validation performance, δ is the improvement threshold, and $\mathbf{1}[\cdot]$ is the indicator function.

Algorithm 2: The workflow of HBS

Input: Batch $\mathcal{B}$, actions (op_{idx}, mag_{idx}), phase p, confidence tracker $\mathcal{C}$, global sample ID i
Parameter: SAC actor π_ϕ, temperature parameter α
Output: Augmented batch $\tilde{\mathcal{B}}$

if $p == 1$ **then**
  Apply uniform augmentation to batches using Eq. (4)
  Update confidence statistics in $\mathcal{C}$ using Eq. (5)
else
  Apply Gaussian noise to states and global ID i
  for each sample j in batch **do**
    Classify sample difficulty using Eq. (6)
    Sample augmentation action using Eq. (7)
    Apply sample-level augmentation $\tilde{x}_j$
  end for
end if
return Augmented batch $\tilde{\mathcal{B}}$

First, the default is to execute the first phase of augmentation. HBS applies uniform augmentation AugOp to the current mini-batch $\mathcal{B} = \{(x_j)\}_{j=1}^{B}$ using selected actions (op_{idx}, mag_{idx}) from the SAC actor network:

$$\tilde{x}_j = \text{AugOp}(x_j, op_{idx}, mag_{idx}), \quad \forall j \in \{1, 2, ..., B\} \tag{4}$$

Meanwhile, HBS tracks individual sample performance using a global sample ID i that uniquely identifies each sample across all training batches. For each sample, HBS maintains statistics $\mathcal{C}_i$:

$$\mathcal{C}_i = \{c_i, g_i, a_i\} \tag{5}$$

where c_i is prediction confidence, g_i is the confidence gap between top-1 and top-2 predictions, and a_i is accuracy rate. These are updated using exponential moving averages with momentum β: $c_i^{new} = \beta \cdot c_i^{old} + (1-\beta) \cdot c_i^{current}$. Statistics are computed only from correct predictions to ensure reliable difficulty assessment.

Once performance saturation is detected via Eq.(3), HBS transitions to Phase 2 for personalized augmentation. Phase 2 assigns different augmentation strategies based on the augmentation difficulty of each sample.

To achieve this personalization, HBS first classifies each sample i into difficulty-aware sets $diff_i$ using the accumulated statistics from Eq.(5):

$$diff_i = \begin{cases} \text{DS} & \text{if } c_i > \tau_1 \wedge g_i > \tau_2 \wedge a_i > \tau_3 \\ \text{AS} & \text{if } c_i > \tau_4 \wedge g_i > \tau_5 \wedge a_i \in [\tau_6, \tau_7] \\ \text{IS} & \text{otherwise} \end{cases} \tag{6}$$

where Discriminative Set (DS) holds well-augmented samples, Ambiguous Set (AS) holds moderately augmented samples, and Indiscriminate Set (IS) holds difficult samples. The thresholds τ_1 to τ_7 are confidence, gap, and accuracy thresholds. During classification, HBS adds Gaussian noise $\mathcal{N}(0, 0.05^2)$ to MDP states in Eq.(2) and perturbs global sample IDs i to prevent overfitting to fixed sample patterns.

Based on the difficulty classification $diff_i$, HBS applies different augmentation strategies to each sample. The augmentation action for the j-th sample in the current batch, with global ID i_j, is determined by its difficulty level $diff_{i_j}$:

$$(op_j, mag_j) = \begin{cases} \pi_\phi(\mathbf{s}_t | \alpha = \alpha_{DS}) & \text{if } diff_{i_j} = \text{DS} \\ \pi_\phi(\mathbf{s}_t | \alpha = \alpha_{AS}) & \text{if } diff_{i_j} = \text{AS} \\ (\text{NoAug}, 0) & \text{if } diff_{i_j} = \text{IS} \end{cases} \tag{7}$$

where α is the temperature parameter controlling exploration intensity. α_{DS} and α_{AS} represent different temperature values for DS and AS sets respectively. The action $(\text{NoAug}, 0)$ indicates no augmentation is applied with zero magnitude level for samples in IS set.

Finally, HBS applies the selected augmentations to each sample according to their individual actions (op_j, mag_j) to generate the augmented spectrograms $\tilde{x}_j$. The augmented batch $\tilde{\mathcal{B}} = \{\tilde{x}_j\}_{j=1}^{B}$ is then returned. Throughout this process, the confidence tracker $\mathcal{C}$ is continuously updated to maintain sample statistics for future difficulty assessment.

Performance evaluation and reward computation. Based on the augmented batch $\tilde{\mathcal{B}}$, we evaluate the efficacy of the augmentation strategies. PASA compares the training benefits of using original versus augmented batch. The reward r_t measures the improvement from augmentation:

$$r_t = \text{BA}(\mathbf{y}_{val}, \hat{y}_{val,aug}) - \text{BA}(\mathbf{y}_{val}, \hat{y}_{val,orig}) \tag{8}$$

where $\mathbf{y}_{val}$ represents ground truth labels on the validation set, $\hat{y}_{val,aug}$ and $\hat{y}_{val,orig}$ are predictions from the diagnostic model f_θ trained separately on augmented batch $\tilde{\mathcal{B}}$ and original batch $\mathcal{B}$ for one step each and then tested on the validation set. BA is balanced accuracy defined as:

$$\text{BA}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c} \tag{9}$$

where C is the number of classes, TP_c is true positives for class c, and FN_c is false negatives for class c.

Afterwards, PASA updates the MDP state to $\mathbf{s}_{t+1}$ using features extracted from both the original batch $\mathcal{B}$ and augmented batch $\tilde{\mathcal{B}}$ according to Eq.(2). This forms the complete experience tuple $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ which is stored in the replay buffer for SAC training.

Network optimization. Finally, with the experience tuple stored, PASA jointly optimizes the diagnostic model and SAC networks with different update frequencies. The diagnostic model f_θ is updated using the augmented batch $\tilde{\mathcal{B}}$ with cross-entropy loss (Mao, Mohri, and Zhong 2023). SAC networks are updated periodically using standard actor-critic loss functions (Haarnoja et al. 2018).

Experiments

Datasets and metrics

We conduct experiments on three respiratory sound datasets including CirCor DigiScope (Oliveira et al. 2022), SPRSound 2022 (Zhang et al. 2022b), and SPRSound 2023 (Zhang et al. 2023). CirCor DigiScope focuses on murmur detection. SPRSound 2022 and 2023 include event-level tasks (T1-1, T1-2) and recording-level tasks (T2-1, T2-2). We use W.acc, UAR (Niizumi et al. 2024a), and Score (%) (Zhang et al. 2022b) as evaluation metrics.

Implementation details

All experiments are conducted using PyTorch on RTX 8000 GPUs with random seed 42. We use ImageNet-pretrained EfficientNet-B4 (Tan and Le 2019) as the diagnostic model, and use AdamW optimizer with a learning rate of 1e-4, a batch size of 32 and an early stopping patience of 20. The SAC network uses a learning rate of 3e-4, a discount factor of 0.99, a target update of 0.005 and a replay buffer size of 5,000. HBS strategy uses confidence threshold $\tau_1 = 0.8, \tau_4 = 0.5$, gap threshold $\tau_2 = 0.5, \tau_5 = 0.1$, accuracy threshold $\tau_3 = 0.7, [\tau_6 = 0.3, \tau_7 = 0.7]$, and improvement threshold $\sigma = 0.001$. Temperature parameters are set to $\alpha_{DS} = 1.0$ and $\alpha_{AS} = 0.3$. All results are averages over five-fold cross-validation. Other details are in Appendix A.

Comparison with state-of-the-art methods

Quantitative results. PASA outperforms SOTA methods on Circor DigiScope dataset for murmur detection, as shown in Table 1. PASA surpasses the results of Niizumi et al. (Niizumi et al. 2024b) with 0.83 W.acc and 0.71 UAR by achieving 0.85 W.acc and 0.73 UAR.

Method	W.acc	UAR
Panah et al. (Panah, Hines, and McKeever 2023)	0.80	0.70
McDonald et al. (McDonald, Gales, and Agarwal 2022)	0.80	0.68
Niizumi et al. (Niizumi et al. 2024b)	0.83	0.71
PASA	**0.85**	**0.73**

Table 1: Quantitative comparison of PASA and other SOTA methods on the Circor DigiScope test dataset.

PASA achieves SOTA results on the SPRSound 2022 dataset for Tasks 1-1 and 2-2. It reaches 90.60% and 64.29%, which is better than Hu et al. (Hu et al. 2025)'s 89.46% and 62.93%. For Tasks 1-2 and 2-1, PASA gets competitive results of 70.39% and 75.71%. These are slightly lower than Hu et al.'s results because multi-class complexity needs diverse data for sample-level personalization. Importantly, PASA achieves these results through pure augmentation improvements without architectural modifications, and still performs well on different tasks. On SPRSound 2023, PASA gets the best results for all four tasks, reaching 84.56%, 70.21%, 78.79%, and 66.07%, with improvements up to 5.28%. Such results show PASA's ability to identify ideal augmentation strategies for various acoustic characteristics.

Qualitative results. Figure 4 shows the performance of PASA on the SPRSound 2022 dataset using confusion matrices and Receiver Operating Characteristic (ROC) curves. PASA handles the hardest mixed condition (CAS & DAS) well, getting 0.806 AUC with only 36 samples. It also performs best in the Poor Quality category for 0.901 AUC. The confusion matrix shows Normal sounds get 67.0% correct classification with few random errors. Other qualitative results are in Appendix B. Most importantly, PASA works better on rare classes that other methods often miss. This demonstrates the superiority of PASA in real medical scenarios, particularly for rare conditions.

Method	T1-1	T1-2	T2-1	T2-2
Zhang et al. (Zhang et al. 2022b)	75.22	61.57	56.71	37.84
Ma et al. (Ma et al. 2022)	84.91	74.73	70.13	52.85
Li et al. (Li et al. 2022)	88.86	82.03	71.79	53.31
Chen et al. (Chen et al. 2022a)	89.00	80.00	71.00	36.00
Chen et al. (Chen et al. 2022b)	89.26	79.70	71.51	-
Zhang et al. (Zhang et al. 2022a)	-	-	71.14	53.14
Babu et al. (Babu et al. 2022)	-	-	-	51.76
Ngo et al. (Ngo et al. 2023b)	84.90	77.40	74.50	53.90
Wang et al. (Wang et al. 2024a),	-	80.18	-	-
Wang et al. (Wang et al. 2024b)	-	-	-	56.99
Hu et al. (Hu et al. 2025)	89.46	**84.03**	**76.35**	62.93
PASA	**90.60**	70.39	75.71	**64.29**

Table 2: Quantitative comparison of our method with SOTA methods on the SPRSound 2022 dataset across four tasks.

Method	T1-1	T1-2	T2-1	T2-2
Babu et al. (Babu et al. 2022)	71.99	59.26	66.53	54.94
Ma et al. (Ma et al. 2022)	73.29	64.60	75.87	53.84
Chen et al. (Chen et al. 2022b)	74.82	59.87	69.86	41.09
Zhang et al. (Zhang et al. 2022a)	81.96	55.51	72.28	52.41
Li et al. (Li et al. 2022)	78.48	64.79	54.71	41.67
Pessoa et al. (Pessoa et al. 2023)	75.60	46.66	65.81	45.83
Ngo et al. (Ngo et al. 2023a)	80.97	66.66	74.43	60.79
Hu et al. (Hu et al. 2025)	76.93	63.18	66.15	51.21
PASA	**84.56**	**70.21**	**78.79**	**66.07**

Table 3: Quantitative comparison of our method with SOTA methods on the SPRSound 2023 dataset across four tasks.

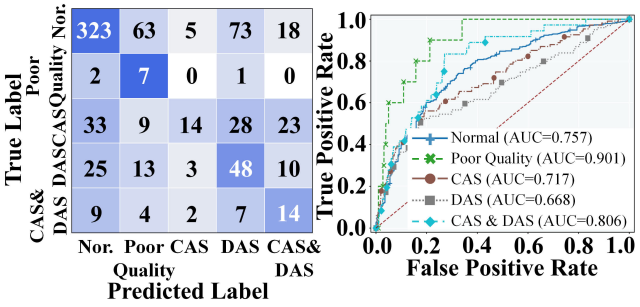

Figure 4: Confusion matrices and ROC curves on the SPRSound 2022 dataset for Task 2-2.

Ablation study

Different components analysis. We test different components of PASA on the SPRSound 2022 dataset across four tasks, as shown in Table 4. The baseline (Base) is EfficientNet-B4, with $O(N)$ complexity. Base+FA (Fixed Augmentation) uses fixed-size transformations. It selects the most frequently used augmentation in PASA and sets its application probability to 0.5. Compared with Base, Base+FA keeps $O(N)$ cost and improves performance by 0.5-2%. Base+BO (Batch-Only) uses batch-level learning, with complexity of $O(N + B \cdot SAC)$. It outperforms Base+FA by 1%, showing that RL is more effective than static augmentation. Base+SO (Sample-Only) uses sample-level learning and the cost is $O(N + N \cdot SAC)$. It performs well but has a

high computational cost. PASA uses HBS strategy with cost $O(N+\alpha \cdot B \cdot SAC+(1-\alpha) \cdot N \cdot SAC)$. Other experiments with individual augmentation of operations and magnitudes are in Appendix C and D. These results show that HBS switches between batch and sample methods smartly and performs better than other methods. PASA not only costs less than Base+SO, but also provides better personalization than Base+BO. This solves the efficiency-personalization trade-off in adaptive augmentation.

Method	T1-1	T1-2	T2-1	T2-2	Time Complexity
Base	89.42	66.18	73.99	62.57	N
Base+FA	90.05	68.80	73.81	62.80	N
Base+BO	90.39	68.70	74.11	63.62	$N + B \cdot S$
Base+SO	90.58	69.35	74.47	63.65	$N + N \cdot S$
PASA	**90.60**	**70.39**	**75.71**	**64.29**	$N + \alpha B \cdot S + (1-\alpha)N \cdot S$

Table 4: Ablation study on the effectiveness of the components on SPRSound 2022 datasets.

Hyperparameter sensitivity analysis. We study parameters that control the HBS strategy behavior of PASA on SPRSound 2022 dataset for Task 2-2. Table 5 shows the effects of gradually adding classification criteria from Eq.(6) and their respective parameters. First, we test only confidence thresholds τ_1 and τ_4 for difficulty classification, finding best values (0.8, 0.5) with 63.85% score. Then we add confidence gap thresholds τ_2 and τ_5, keeping the best confidence values and testing gap combinations. The best gap setting is (0.5, 0.1). Next, we add accuracy thresholds τ_3 for DS and range $[\tau_6, \tau_7]$ for AS. The final best combination is 0.7 and [0.3, 0.7], giving 64.29% score. This experiment proves that each of the factors in the difficulty classification criteria makes results better. Moreover, we test temperature parameters from Eq.(7) to control the action space under different difficulty-aware sets. The results show that $\alpha_{DS} = 1.0$ and $\alpha_{AS} = 0.3$ are optimal for different sample difficulty levels.

Parameter Group	Value	T2-2
Confidence Only τ_1, τ_4	(0.6, 0.3)	63.21
	(0.7, 0.4)	63.57
	(0.8, 0.5)	**63.85**
	(0.9, 0.6)	63.42
+ Confidence Gap τ_2, τ_5	(0.1, 0.05)	63.91
	(0.3, 0.08)	63.98
	(0.5, 0.1)	**64.08**
	(0.7, 0.15)	63.94
+ Accuracy τ_3, $[\tau_6, \tau_7]$	0.5, [0.2, 0.5]	63.95
	0.6, [0.25, 0.6]	64.11
	0.7, [0.3, 0.7]	**64.29**
	0.8, [0.35, 0.8]	64.15
Temperature α_{DS}, α_{AS}	(0.8, 0.24)	64.01
	(1.0, 0.3)	**64.29**
	(1.2, 0.36)	64.15
	(1.5, 0.45)	63.87

Table 5: Hyperparameter sensitivity analysis on SPRSound 2022 dataset for Task 2-2.

Learning dynamics analysis. Figure 5 shows the process of PASA learning on the SPRSound 2022 dataset for Task 2-2. (a) The key transition happens at epoch 28. PASA uses sample-level augmentation to avoid overfitting. While Base+BO and Base+SO methods show growing gaps between training and validation accuracy, PASA keeps stable validation performance. (b) Sample difficulty changes over time. At first, most samples are in IS because there isn't enough data for learning. As the model learns, DS samples increase while AS samples stay stable. This allows different augmentations for different samples. (c) Augmentation preferences change during training. Early stages show balanced preferences for all operations. Later stages prefer harmonic perturbations and spectral contrast while avoiding harmful operations. Other preferences of tasks are in Appendix E. These results show PASA can find optimal strategies and achieve effective learning via HBS.

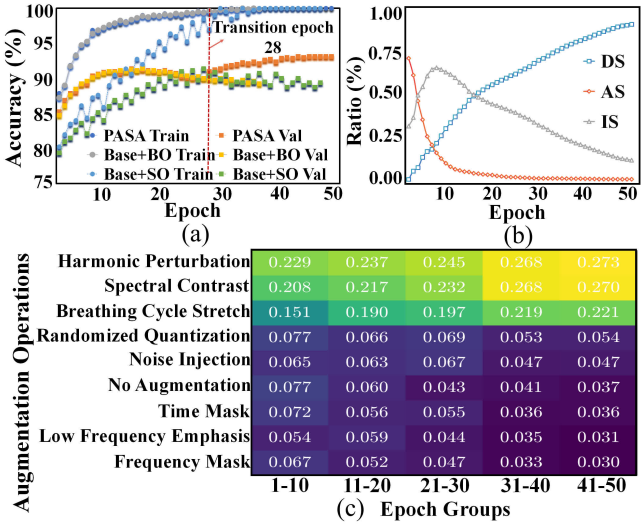

Figure 5: Learning dynamics analysis on SPRSound 2022 dataset for Task 2-2. (a) Accuracy comparison among Base+BO, Base+SO, and PASA on training and validation sets, respectively. (b) Evolution of difficulty-aware sets. (c) Augmentation operation usage across epoch groups.

Conclusions

We propose a PASA, the first RL-based adaptive augmentation for automated auscultation, achieving 12.6% improvement over SOTA on SPRSound 2023. The HBS strategy balances efficiency and personalization with intelligent phase transitions. Notably, our findings reveal that pathological sounds require personalized augmentation strategies, which challenge the use of uniform approaches in medical AI. This paradigm shift is beneficial for data-scarce medical domains, as adaptive learning maximizes limited resources. Future work will extend PASA to multimodal applications and other medical domains.

Acknowledgments

This work was supported by the Macao Polytechnic University under grant RP/FCA-04/2024.

References

Abayomi-Alli, O. O.; Damaševičius, R.; Qazi, A.; Adedoyin-Olowe, M.; and Misra, S. 2022. Data augmentation and deep learning methods in sound classification: A systematic review. *Electronics*, 11(22): 3795.

Al-Hasan, T. M.; Noorizadeh, M.; Bensaali, F.; Meskin, N.; and Hssain, A. A. 2025. Current trends and future orientation in diagnosing lung pathologies: A systematic survey. *Intelligent Medicine*, 05(01): 23–36.

Babu, N.; Kumari, J.; Mathew, J.; Satija, U.; and Mondal, A. 2022. Multiclass Categorisation of Respiratory Sound Signals Using Neural Network. In *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 228–232.

Bernardi, S.; Giudici, F.; Leone, M. F.; Zuolo, G.; Furlotti, S.; Carretta, R.; and Fabris, B. 2019. A prospective study on the efficacy of patient simulation in heart and lung auscultation. *BMC medical education*, 19: 1–7.

Chen, Z.; Wang, H.; Yeh, C.-H.; and Liu, X. 2022a. Classify respiratory abnormality in lung sounds using stft and a fine-tuned resnet18 network. In *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 233–237. IEEE.

Chen, Z.; Wang, H.; Yeh, C.-H.; and Liu, X. 2022b. Classify Respiratory Abnormality in Lung Sounds Using STFT and a Fine-Tuned ResNet18 Network. In *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 233–237.

Fawzi, A.; Samulowitz, H.; Turaga, D.; and Frossard, P. 2016. Adaptive data augmentation for image classification. *2016 IEEE International Conference on Image Processing (ICIP)*, 3688–3692.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning*, 1861–1870. PMLR.

Heitmann, J.; Glangetas, A.; Doenz, J.; Dervaux, J.; Shama, D. M.; Garcia, D. H.; Benissa, M. R.; Cantais, A.; Perez, A.; Müller, D.; et al. 2023. DeepBreath—automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries. *NPJ digital medicine*, 6(1): 104.

Hou, C.; Zhang, J.; and Zhou, T. 2023. When to learn what: Model-adaptive data augmentation curriculum. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1717–1728.

Hu, J.; Leow, C. S.; Tao, S.; Goh, W. L.; and Gao, Y. 2025. Supervised Contrastive Learning Framework and Hardware Implementation of Learned ResNet for Real-Time Respiratory Sound Classification. *IEEE Transactions on Biomedical Circuits and Systems*, 19(1): 185–195.

Kim, J.-W.; Yoon, C.; Toikkanen, M.; Bae, S.; and Jung, H.-Y. 2023. Adversarial Fine-tuning using Generated Respiratory Sound to Address Class Imbalance. *arXiv preprint arXiv:2311.06480*.

Li, J.; Wang, J.; Chen, J.; and Xu, T. 2025. Towards Robust Point Cloud Recognition With Sample-Adaptive Auto-Augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 3003–3017.

Li, J.; Wang, X.; Wang, X.; Qiao, S.; and Zhou, Y. 2022. Improving The ResNet-based Respiratory Sound Classification Systems With Focal Loss. In *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 223–227.

Liu, H.-T. D.; Williams, F.; Jacobson, A.; Fidler, S.; and Litany, O. 2022. Learning smooth neural functions via lipschitz regularization. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–13.

Ma, W.-B.; Deng, X.-Y.; Yang, Y.; and Fang, W.-C. 2022. An Effective Lung Sound Classification System for Respiratory Disease Diagnosis Using DenseNet CNN Model with Sound Pre-processing Engine. In *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 218–222.

Mao, A.; Mohri, M.; and Zhong, Y. 2023. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, 23803–23828. pmlr.

McDonald, A.; Gales, M. J.; and Agarwal, A. 2022. Detection of Heart Murmurs in Phonocardiograms with Parallel Hidden Semi-Markov Models. In *2022 Computing in Cardiology (CinC)*, volume 498, 1–4.

Müller, S. G.; and Hutter, F. 2021. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 774–782.

Ngo, D.; Pham, L.; Phan, H.; Tran, M.; and Jarchi, D. 2023a. A Deep Learning Architecture with Spatio-Temporal Focusing for Detecting Respiratory Anomalies. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 1–5. IEEE.

Ngo, D.; Pham, L.; Phan, H.; Tran, M.; Jarchi, D.; and Kolozali, Ş. 2023b. An inception-residual-based architecture with multi-objective loss for detecting respiratory anomalies. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, 1–6. IEEE.

Niizumi, D.; Takeuchi, D.; Ohishi, Y.; Harada, N.; and Kashino, K. 2024a. Exploring Pre-trained General-purpose Audio Representations for Heart Murmur Detection. *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4.

Niizumi, D.; Takeuchi, D.; Ohishi, Y.; Harada, N.; and Kashino, K. 2024b. Exploring Pre-trained General-purpose Audio Representations for Heart Murmur Detection. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4. IEEE.

Oliveira, J.; Renna, F.; Costa, P. D.; Nogueira, M.; Oliveira, C.; Ferreira, C.; Jorge, A.; Mattos, S.; Hatem, T.; Tavares, T.; Elola, A.; Rad, A. B.; Sameni, R.; Clifford, G. D.; and Coimbra, M. T. 2022. The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification. *IEEE Journal of Biomedical and Health Informatics*, 26(6): 2524–2535.

Osama, S.; Salah, L.; Dahi, G.; Ghazal, M.; Almajali, E.; Hussain, A. J.; Yousaf, J.; and Hassan, T. 2024. A Review of Screening Heart and Lung Diseases using Auscultation and Artificial Intelligence. In *2024 17th International Conference on Development in eSystem Engineering (DeSE)*, 168–173. IEEE.

Panah, D. S.; Hines, A.; and McKeever, S. 2023. Exploring Wav2vec 2.0 Model for Heart Murmur Detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*, 1010–1014.

Park, D. S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E. D.; and Le, Q. V. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech*.

Park, J. S.; Park, S.-Y.; Moon, J. W.; Kim, K.; and Suh, D. I. 2025. Artificial Intelligence Models for Pediatric Lung Sound Analysis: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, 27: e66491.

Pessoa, D.; Petmezas, G.; Papageorgiou, V. E.; Rocha, B. M.; Stefanopoulos, L.; Kilintzis, V.; Maglaveras, N.; Frerichs, I.; de Carvalho, P.; and Paiva, R. P. 2023. Pediatric Respiratory Sound Classification Using a Dual Input Deep Learning Architecture. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 1–5. IEEE.

Ren, Z.; Chang, Y.; Nguyen, T. T.; Tan, Y.; Qian, K.; and Schuller, B. W. 2024. A comprehensive survey on heart sound analysis in the deep learning era. *IEEE Computational Intelligence Magazine*, 19(3): 42–57.

Sun, Y.; Xu, K.; Liu, C.; Dou, Y.; Wang, H.; Ding, B.; and Pan, Q. 2024. Automated data augmentation for audio classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 2716–2728.

Sutton, R.; and Barto, A. 1998. Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks*, 9(5): 1054–1054.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 6105–6114. PMLR.

Wang, F.; Yuan, X.; Bao, J.; Lam, C.-T.; Huang, G.; and Chen, H. 2024a. OFGST-Swin: Swin transformer utilizing overlap fusion-based generalized S-transform for respiratory cycle classification. *IEEE Transactions on Instrumentation and Measurement*.

Wang, F.; Yuan, X.; Liu, Y.; and Lam, C.-T. 2024b. LungNeXt: A novel lightweight network utilizing enhanced mel-spectrogram for lung sound classification. *Journal of King Saud University-Computer and Information Sciences*, 36(8): 102200.

Wang, Z.; and Wang, Z. 2022. 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)*, 9017–9021.

Whang, S. E.; Roh, Y.; Song, H.; and Lee, J.-G. 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 32(4): 791–813.

Wu, H.; Lei, C.; Sun, X.; Wang, P.-S.; Chen, Q.; Cheng, K.-T.; Lin, S.; and Wu, Z. 2023. Randomized quantization: A generic augmentation for data agnostic self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16305–16316.

Yu, X.; Tian, Y.; Wang, L.; Feng, P.; Wu, W.; and Shi, R. 2024. AdaptAUG: Adaptive Data Augmentation Framework for Multi-Agent Reinforcement Learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 10814–10820.

Zhang, L.; Zhu, Y.; Tu, S.; and Xu, L. 2022a. A Feature Polymerized Based Two-Level Ensemble Model for Respiratory Sound Classification. In *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 238–242.

Zhang, Q.; Zhang, J.; Yuan, J.; Huang, H.; Zhang, Y.; Chen, C.; Lin, J.; Zhang, B.; Lv, G.; Lin, S.; Wang, N.; Liu, X.; Tang, M.; Wang, Y.; Liu, L.; Ma, H.; Xie, D.; Wu, L.; Yang, H.; Yuan, S.; Chen, M.; Zhang, B.; Zhou, H.; Zhao, J.; Li, Y.; Yin, Y.; Zhao, L.; Wang, G.; and Lian, Y. 2023. Grand Challenge on Respiratory Sound Classification for SPRSound Dataset. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 1–5.

Zhang, Q.; Zhang, J.; Yuan, J.; Huang, H.; Zhang, Y.; Zhang, B.; Lv, G.; Lin, S.; Wang, N.; Liu, X.; Tang, M.; Wang, Y.; Ma, H.; Liu, L.; Yuan, S.; Zhou, H.; Zhao, J.; Li, Y.; Yin, Y.; Zhao, L.; Wang, G.; and Lian, Y. 2022b. SPRSound: Open-Source SJTU Paediatric Respiratory Sound Database. *IEEE Transactions on Biomedical Circuits and Systems*, 16(5): 867–881.

Zhao, Q.; Geng, S.; Wang, B.; Sun, Y.; Nie, W.; Bai, B.; Yu, C.; Zhang, F.; Tang, G.; Zhang, D.; Zhou, Y.; Liu, J.; and Hong, S. 2024. Deep Learning in Heart Sound Analysis: From Techniques to Clinical Applications. *Health Data Science*, 4: 0182.