forum acusticum 2026
8-12 September · Graz, Austria

SPSC-HCM-16C: A MULTI-CHANNEL LUNG SOUND BENCHMARK

Minh D. N. Nguyen[1], Truc T. K. Nguyen[2], Si V. Nguyen[3,4], Thao T. N. Tran[4], Nghia T. H. Ma[4], Hung T. Quach[4], An L. Pham[3,4], Benedikt Mayrhofer[5], Martin Hagmüller[5], Franz Pernkopf[5]

[1] *FPT University, Da Nang, Vietnam*
[2] *The University of Da Nang - University of Science and Technology, Da Nang, Vietnam*
[3] *University of Medicine and Pharmacy at Ho Chi Minh City, Ho Chi Minh City, Vietnam*
[4] *Nguyen Trai Hospital, Ho Chi Minh City, Vietnam*
[5] *Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

This paper presents *SPSC-HCM-16C*, a benchmark dataset[1] for patient-level respiratory disease classification from synchronized 16-channel lung sound recordings collected in a real clinical environment. The dataset contains recordings from 183 subjects, including healthy controls and four respiratory disease groups, and supports three diagnostic settings: 2-class, 3-class, and 5-class classification. The aim is to establish a standardized and reproducible benchmark for future research in multi-channel computational auscultation. To this end, we define subject-independent evaluation protocols, including a fixed 60/20/20 split and an 80/20 split with 5-fold cross-validation on the training-validation subset, together with a transparent reference baseline. The baseline uses MFCC features, a modified 16-channel ResNet101 backbone with partial fine-tuning of Layer3, and mean pooling for patient-level aggregation. Under the fixed 60/20/20 split, the baseline achieves Macro-F1 scores of 93.68%, 77.06%, and 44.61% for the 2-class, 3-class, and 5-class tasks, respectively, with consistent trends observed under 5-fold cross-validation. These results establish a reproducible reference for future comparison and highlight the challenges posed by weak supervision, class imbalance, and inter-class similarity in multi-channel lung sound analysis.

***Corresponding author**: ntktruc@dut.udn.vn.*

[1]The dataset and benchmark algorithms are available at [1]. Ethical approval was granted by the Ethics Committee of the University of Medicine and Pharmacy Socialist Republic of Vietnam at Ho Chi Minh City (Reference No. 1241/HĐĐĐ-ĐHYD).

Keywords: *lung sound dataset, multi-channel auscultation, patient-level diagnosis, respiratory disease classification*

1. Introduction

Lung auscultation is a widely used non-invasive technique for respiratory assessment, but its interpretation remains subjective and affected by inter-observer variability, environmental noise, and heterogeneous sound patterns across patients and recording sites [2], [3].

Although automated lung sound analysis has progressed rapidly, reproducible comparison across studies remains difficult. Many existing works rely on different datasets, preprocessing pipelines, label definitions, and evaluation protocols, making reported performance only partially comparable across papers [3], [4], [5]. Public resources such as the ICBHI 2017 respiratory sound database [4] and the HF_Lung_V1 dataset [5] have played an important role in advancing the field, yet most widely used benchmarks are based on single-channel recordings and are primarily designed for event detection or segment-level classification rather than patient-level diagnosis. As a result, they only partially reflect the clinical setting in which diagnostic decisions are made for the patient as a whole.

Another limitation is that clinically relevant respiratory evidence is often spatially distributed across the chest and temporally sparse within a recording. This makes multi-channel auscultation particularly appealing, because synchronized acquisition from multiple chest locations can preserve spatial acoustic information that is unavailable in conventional single-channel recordings. Prior studies have shown that multi-channel learning can improve respiratory sound classification by jointly exploiting spectral, temporal, and spatial cues [6], [7]. At the same time, real-world respiratory datasets commonly provide labels only at

the patient-level, while fine-grained temporal annotations of abnormal events are rarely available. Unlike our prior model-centric research using this data [8], the present paper focuses on benchmark standardization, including dataset documentation, subject-independent protocols, and a transparent baseline reference. In this work, we provide SPSC-HCM-16C as a multi-channel lung sound benchmark dataset including task definitions, subject-independent evaluation protocols, and a reproducible reference baseline.

2. Dataset and Benchmark Setup

2.1. Clinical Acquisition and Study Population

The SPSC-HCM-16C dataset was collected at Nguyen Trai Hospital, Ho Chi Minh City, Vietnam, during routine respiratory examinations of adult subjects (≥18 years old). Lung sounds were acquired using a synchronized 16-channel electronic stethoscope developed at Graz University of Technology, Austria. The acquisition system (see Figure 1) enables simultaneous recording from multiple auscultation sites on the chest, thereby preserving spatially aligned acoustic information that is not available in conventional single-channel recordings.

All recordings were performed following standard clinical auscultation procedures in a real hospital environment. Diagnostic labels were assigned at the patient level based on clinical assessment and medical records by experienced physicians. In line with routine clinical practice, no fine-grained temporal annotations of abnormal respiratory events were available. The dataset therefore reflects a weakly supervised setting in which only patient-level diagnostic labels are provided.

The dataset consists of recordings from 183 subjects, including 92 healthy controls and 91 patients with respiratory diseases. The pathological cases cover four disease categories: *Asthma*, *COPD*, *Pleural Effusion*, and *Pneumonia*. Each subject contributed one or more recordings, with an average duration of approximately 40 seconds. Each recording contains *16 synchronized audio channels* captured concurrently from different auscultation locations. This design preserves inter-site temporal alignment and supports patient-level multi-channel respiratory sound analysis. An overview of subject demographics and class distribution is summarized in Table 1.

In addition to the synchronized lung sound recordings, chest X-ray images were collected for subjects in the disease group to support diagnostic verification, but not for healthy controls. Therefore, although auxiliary imaging data are available for some pathological cases, the present benchmark is intentionally restricted to the audio modality, and all reported baseline results are based on synchronized multi-channel lung sound recordings only.

2.2. Benchmark Tasks and Protocols

To support evaluation under different diagnostic granularities, SPSC-HCM-16C is organized into three tasks:

- **2-class classification:** Healthy vs. Unhealthy,
- **3-class classification:** Healthy, Chronic (Asthma, COPD), and Non-chronic (Pneumonia, Pleural Effusion),
- **5-class classification:** Healthy, Asthma, COPD, Pneumonia, and Pleural Effusion.

Figure 1: Sixteen-channel electronic stethoscope used for multi-channel lung sound acquisition in the clinical setting.

Table 1: Overview of subject demographics and class distribution in the SPSC-HCM-16C dataset.

Class	Subj.	Male	Female	Avg. Age
Asthma	14	1	13	65.00
COPD	23	23	0	69.35
Pleural Effusion	20	14	6	59.65
Pneumonia	34	20	14	66.85
Control/Healthy	92	63	29	42.93

The benchmark defines two subject-independent evaluation protocols. The first uses a fixed *60/20/20* partition of subjects into training, validation, and test subsets. The second uses an *80/20* subject-level split, where 80% of the subjects are assigned to a training-validation pool and 20% are reserved as an independent test set; a *5-fold cross-validation* procedure is then performed only on the training-validation subset. In all settings, strict *subject independence* is enforced so that recordings or segments from the same subject never appear in both training and evaluation subsets.

From an application perspective, the three task settings reflect different levels of clinical utility. The 2-class task is relevant to broad screening, the 3-class task supports coarse disease grouping, and the 5-class task serves as a more stringent benchmark for fine-grained respiratory disease discrimination.

2.3. Standardized Preprocessing

To ensure comparability across benchmark results, all experiments follow the same preprocessing pipeline. Each lung sound recording is resampled to 4 kHz. The signal from each channel is segmented into fixed windows of 6 seconds with 50% overlap. For each segment, 40-dimensional MFCC features are extracted independently from each of the 16 synchronized channels. The channel-wise features are then stacked into

a multi-channel time-frequency representation.
Although segmentation and feature extraction are performed at the segment level, all metadata remain linked to the corresponding patient identity, class label, and benchmark split. This supports consistent subject-level grouping during training and evaluation and reflects the patient-level decision setting targeted by the benchmark.
To support reproducible benchmarking, the resource is organized around subject-level metadata, split definitions, preprocessing specifications, and evaluation settings corresponding to the two benchmark protocols. This ensures consistency across future studies using the same patient-level benchmark setup. The anonymized dataset and the baseline code are available at [1]. Ethical approval for recording the data was granted by the Ethics Committee of the University of Medicine and Pharmacy Socialist Republic of Vietnam at Ho Chi Minh City (Reference No. 1241/HĐĐĐ-ĐHYD).

3. Baseline Reference Pipeline

The baseline is designed as a simple, transparent, and reproducible reference model for benchmarking on SPSC-HCM-16C. Its purpose is not to maximize performance, but to provide a stable patient-level pipeline that can serve as a common reference for future comparison.
Each patient is represented as a bag of segments. In the configuration, a fixed number of $K = 20$ segments is sampled from the available segment pool of that subject; if fewer than 20 segments are available, sampling with replacement is applied. Each sampled segment is represented by a multi-channel MFCC tensor of size $16 \times F \times T$ and is first normalized using cepstral mean and variance normalization (CMVN).
The baseline adopts a single-encoder architecture based on ResNet101 (see Figure 2). To accommodate the multi-channel MFCC input, the first convolutional layer is modified to accept 16 input channels and produce 64 feature maps. The normalized segment tensor is then processed by the ResNet101 backbone, where only *Layer3*, corresponding to the stage with 23 bottleneck residual blocks, is unfrozen for partial fine-tuning, while all remaining backbone stages are kept fixed. After global average pooling, the backbone output is flattened into a 2048-dimensional feature vector and projected to a 256-dimensional segment embedding through a fully connected layer.
Let $\mathbf{h}_i \in \mathbb{R}^{256}$ denote the embedding of the i-th segment in a patient bag. The patient-level representation is obtained by mean pooling over the K segment embeddings:

$$\mathbf{z} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{h}_i. \tag{1}$$

The pooled feature vector $\mathbf{z} \in \mathbb{R}^{256}$ is then passed to a final linear classifier with C output units, where $C \in \{2, 3, 5\}$ depending on the target task. In this way, the same backbone and aggregation pipeline are used for the 2-class, 3-class, and 5-class settings, with only the output dimension of the classifier adjusted according to the label space.
Training is performed using AdamW with patient-level weighted sampling to alleviate class imbalance, together with early stopping based on validation performance. For reproducibility, the loss function and any applied data augmentation are explicitly indicated in the result table. Unless otherwise stated, the reference configuration uses the ResNet101 single-encoder baseline with the modified 16-channel Conv1 layer, only Layer3 is unfrozen for partial fine-tuning, mean pooling over $K = 20$ segment embeddings, and a task-specific linear classifier. The role of this baseline is to provide a transparent and computationally accessible reference for future comparison [8].

4. Results and Benchmark Analysis

All experiments were conducted under the two benchmark protocols described above. The baseline was used without task-specific architectural modifications beyond the final output layer, and all predictions were evaluated at the *patient level*. To assess result stability and reduce the influence of stochastic training effects, each experimental setting was independently repeated three times using different random seeds. All reported values are presented as mean ± standard deviation over the three runs.
For the 2-class task, we report precision (P^+), sensitivity (Se), specificity (Sp), and F1 score. For the 3-class and 5-class tasks, Macro-F1 is used as the primary metric, while macro-averaged precision, sensitivity, and specificity are additionally reported for completeness. Macro-F1 is particularly suitable for SPSC-HCM-16C because it gives equal importance to all classes and is therefore more informative than micro-averaged measures under class imbalance, especially for minority disease categories that are clinically important.
For the binary task, the evaluation metrics are derived from the patient-level confusion matrix:

$$P^+ = \frac{TP}{TP + FP}, \mathrm{Se} = \frac{TP}{TP + FN}, \mathrm{Sp} = \frac{TN}{TN + FP},$$

where TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative counts, respectively. The corresponding F1-score is given by

$$\mathrm{F1} = \frac{2P^+\mathrm{Se}}{P^+ + \mathrm{Se}}.$$

For the multi-class tasks, these quantities are computed in a one-vs-rest manner for each class and then averaged across classes. The Macro-F1 score is defined as

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{F1}_c,$$

where C denotes the number of classes and $\mathrm{F1}_c$ is the class-wise F1-score for class c.

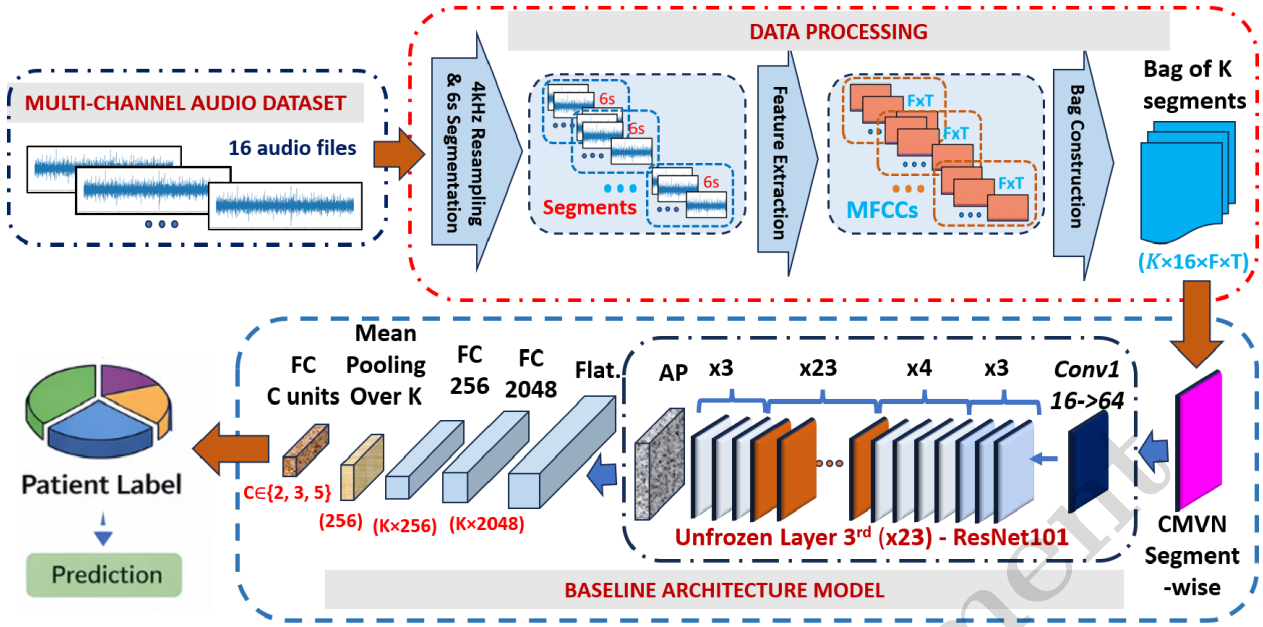

Figure 2: Baseline framework. Multi-channel lung sound recordings are resampled, segmented, converted into MFCC representations, and grouped into patient-level bags. After CMVN normalization, each segment is processed by a modified Conv1 layer and a ResNet101 backbone, where only Layer3 (23 bottleneck blocks) is unfrozen for partial fine-tuning. The backbone output is globally pooled and flattened into a 2048-dimensional feature vector, projected to 256 dimensions, aggregated by mean pooling over K segments, and finally classified at patient-level for the 2-class, 3-class, and 5-class tasks.

4.1. Training Settings

For reproducibility, the main hyperparameters used in the reported experiments are summarized as follows:

- **Backbone and fine-tuning:** ResNet101 with a modified Conv1 layer for 16-channel input; only Layer3 (23 bottleneck blocks) is unfrozen, while all remaining backbone stages are kept fixed.
- **Patient representation:** Each patient is represented by a bag of $K = 20$ segments, and each segment is projected to a 256-dimensional embedding.
- **Optimization:** AdamW optimizer with learning rates of 1×10^{-4} for the 2-class task and 2×10^{-4} for the 3-class and 5-class tasks, weight decay of 1×10^{-4}, batch size of 8, and a maximum of 40 training epochs.
- **Stabilization:** Early stopping with a patience of 8 epochs, gradient clipping with maximum norm 2.0, and patient-level weighted sampling to alleviate class imbalance.

The reported training configurations differ in terms of loss function and optional data augmentation:

- **Poly-2:** Baseline trained with second-order polynomial loss [9].
- **Focal:** Baseline trained with focal loss [10].
- **CyclicShift + GaussianNoise + Poly-2:** Baseline trained with Poly-2 loss together with temporal cyclic shift and additive Gaussian noise, which are classical label-preserving perturbations for audio-like inputs [11].
- **SpecAug + Mixup + Poly-2:** Baseline trained with Poly-2 loss together with spectrogram masking and interpolation-based augmentation [12], [13].

Table 2 is intended as a compact comparison of candidate training settings for the baseline. To limit experimental effort, these values are reported from single runs and are used only to assess the relative effect of different loss and augmentation strategies. The strongest configuration for each task and protocol was subsequently re-evaluated over three independent runs in Table 3.

Table 3 summarizes the best baseline results across the three tasks under the two benchmark protocols. Several consistent observations can be made. First, the baseline achieves the strongest performance in the 2-class setting under both protocols, confirming that broad healthy-versus-unhealthy screening is the easiest task on SPSC-HCM-16C. Second, performance decreases noticeably as the label space becomes more fine-grained, with Macro-F1 dropping from 2-class to 3-class and further to 5-class classification. This trend indicates that benchmark difficulty increases substantially when the model is required to distinguish among clinically related respiratory diseases rather

Table 2: Comparison of training settings for the baseline across tasks and benchmark protocols, reported in terms of Macro-F1 (%). These values correspond to single-run results and are provided to illustrate the relative effect of different loss and augmentation settings.

Task	Protocol	Poly-2	Focal	CyclicShift + GaussianNoise + Poly-2	SpecAug + Mixup + Poly-2
2-class	Fixed 60/20/20	**93.68**	90.08	87.31	90.04
2-class	80/20 + 5-fold CV	89.12	**89.18**	89.12	83.77
3-class	Fixed 60/20/20	**77.06**	62.36	75.28	72.14
3-class	80/20 + 5-fold CV	55.85	65.57	**66.80**	65.94
5-class	Fixed 60/20/20	39.84	43.54	**44.61**	37.16
5-class	80/20 + 5-fold CV	32.27	36.49	**47.09**	37.97

Table 3: Best patient-level benchmark results across the three tasks under the two evaluation protocols on SPSC-HCM-16C. Metrics are reported in %. For the selected best-performing setting of each task and protocol, results are shown as mean ± standard deviation over three independent runs.

Task	Protocol	Training setting	P^+	Se	Sp	Macro-F1
2-class	Fixed 60/20/20	ResNet101 + Poly-2	91.53 ± 7.50	96.30 ± 6.42	91.23 ± 8.04	$\mathbf{93.68 \pm 5.63}$
2-class	80/20 + 5-fold CV	ResNet101 + Focal	87.28 ± 4.63	86.12 ± 3.45	87.59 ± 5.21	$\mathbf{86.28 \pm 4.63}$
3-class	Fixed 60/20/20	ResNet101 + Poly-2	79.35 ± 8.76	76.72 ± 4.89	90.28 ± 2.89	$\mathbf{77.06 \pm 5.57}$
3-class	80/20 + 5-fold CV	ResNet101 + CyclicShift + GaussianNoise + Poly-2	75.50 ± 7.21	67.61 ± 6.61	87.02 ± 2.88	$\mathbf{68.02 \pm 7.34}$
5-class	Fixed 60/20/20	ResNet101 + CyclicShift + GaussianNoise + Poly-2	45.87 ± 9.66	47.76 ± 2.80	91.09 ± 0.92	$\mathbf{44.61 \pm 4.51}$
5-class	80/20 + 5-fold CV	ResNet101 + CyclicShift + GaussianNoise + Poly-2	41.44 ± 12.42	35.87 ± 8.41	88.27 ± 1.69	$\mathbf{37.89 \pm 7.25}$

than only separate healthy from pathological cases.

Fig. 3 highlights the main benchmark trend across the two evaluation protocols. In both cases, the baseline achieves the highest Macro-F1 in the 2-class setting, followed by the 3-class and 5-class settings. This monotonic decrease indicates that SPSC-HCM-16C supports increasingly difficult patient-level classification tasks as label granularity increases.

Fig. 4 provides a more detailed view of the classification behavior across the three benchmark tasks. In the 2-class setting, the confusion matrix shows minimal misclassification, indicating strong separability between healthy and pathological cases. In the 3-class setting, the dominant errors occur between the chronic and non-chronic categories, reflecting partial overlap in respiratory acoustic patterns. In the 5-class setting, the confusion matrix reveals substantially greater ambiguity, with frequent confusion between clinically related conditions such as pneumonia and COPD, and between pleural effusion and pneumonia. Asthma is distributed across multiple predicted classes, suggesting weaker class-specific acoustic signatures in the most fine-grained setting.

From a clinical perspective, the observed confusion patterns are plausible rather than arbitrary. In particular, confusion between chronic and non-chronic categories in the 3-class setting, and among pneumonia, pleural effusion, and COPD in the 5-class setting, suggests that the benchmark captures realistic overlap in respiratory acoustic manifestations. This property is desirable for benchmark design because it avoids artificially easy separation and better reflects practical diagnostic uncertainty.

5. Conclusion

This paper presented *SPSC-HCM-16C*, a benchmark for patient-level respiratory disease classification from synchronized 16-channel lung sound recordings. Rather than proposing a new high-performance architecture, this work focuses on establishing a standardized and reproducible evaluation framework for three tasks (2-class, 3-class, and 5-class), two subject-independent benchmark protocols and a transparent reference baseline based on multi-channel MFCC features, a modified 16-channel ResNet101 backbone with partial fine-tuning of Layer3, and mean pooling. Under the fixed 60/20/20 protocol, the baseline achieves Macro-F1 scores of 93.68%, 77.06%, and 44.61% for the 2-class, 3-class, and 5-class tasks, respectively, while showing the same overall trend under cross-validation. These results indicate that SPSC-HCM-16C supports a useful spectrum, ranging from broad screening to fine-grained disease recognition. Overall, SPSC-HCM-16C provides a resource for systematic and fair comparison of future patient-level respiratory sound analysis methods under standardized evaluation protocols.

References

[1] M. D. N. Nguyen et al., "SPSC-HCM-16C: A multi-channel lung sound benchmark," in *Technical Report, Graz University of Technology*, 2026. DOI: 10.3217/36dmb-qdb49.

[2] H. Pasterkamp, S. S. Kraman, and G. R. Wodicka, "Respiratory sounds: Advances beyond the stethoscope," *American Journal of Respiratory and Critical Care Medicine*, vol. 156, no. 3, pp. 974–987, 1997. DOI: 10.1164/ajrccm.156.3.9701115.

[3] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PLOS ONE*, vol. 12, no. 5, e0177926, 2017. DOI: 10.1371/journal.pone.0177926.

[4] B. M. Rocha et al., "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological Measurement*, vol. 40,

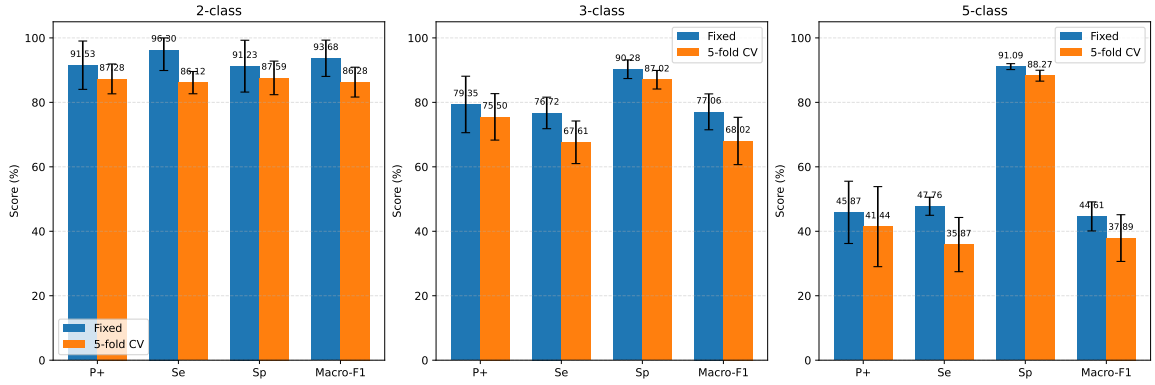

Figure 3: Comparison of patient-level benchmark metrics between the fixed 60/20/20 split and the 80/20 + 5-fold cross-validation protocol across the 2-class, 3-class, and 5-class tasks.

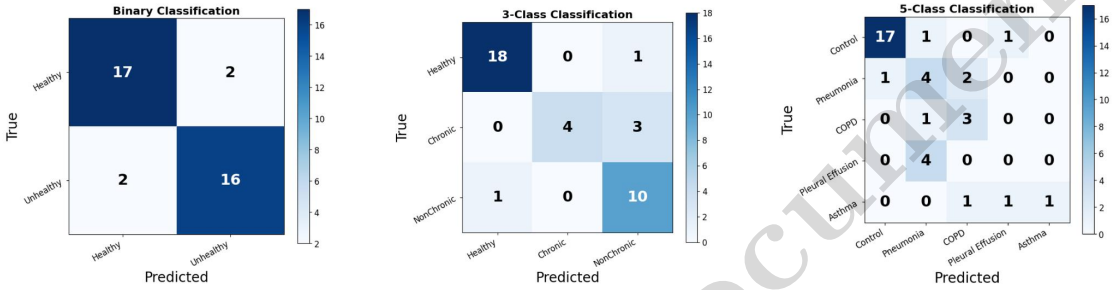

Figure 4: Confusion matrices for the three benchmark tasks under the fixed 60/20/20 split. The results show strong separability in the binary setting, moderate confusion between chronic and non-chronic classes in the 3-class setting, and substantially increased ambiguity in the 5-class setting.

no. 3, p. 035 001, 2019. DOI: 10.1088/1361-6579/ab03ea.

[5] F.-S. Hsu, S.-R. Huang, C.-W. Huang, C.-J. Huang, Y.-R. Cheng, C.-C. Chen, et al., "Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database—hf_lung_v1," *PLOS ONE*, vol. 16, no. 7, e0254134, 2021. DOI: 10.1371/journal.pone.0254134.

[6] E. Messner et al., "Multi-channel lung sound classification with convolutional recurrent neural networks," *Computers in Biology and Medicine*, vol. 122, p. 103 831, 2020. DOI: 10.1016/j.compbiomed.2020.103831.

[7] 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. DOI: 10.1109/TBME.2022.3157514.

[8] T. T. K. Nguyen et al., "Lung-sound respiratory disease classification via multiple-instance learning," *IEEE Access*, 2026. DOI: 10.1109/ACCESS.2026.3669914.

[9] J. Leng, Y. Liu, T. Sun, Y. Wang, and W. Zhang, "Polyloss: A polynomial expansion perspective of classification loss functions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 787–797.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[11] T. Ko, V. Peddinti, D. Povey, and M. L. Seltzer, "Audio augmentation for speech recognition," in *Proceedings of Interspeech*, 2015, pp. 3586–3589.

[12] D. S. Park et al., "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of Interspeech*, 2019, pp. 2613–2617.

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