Semi-Supervised Pulmonary Auscultation Analysis with Cross Pseudo Supervision

Jieruei Chang
Princeton High School
Princeton, New Jersey, USA
jierueic@gmail.com

Ching-Chun Huang
Department of Computer Science
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
chingchun@nycu.edu.tw

***Abstract*—Pulmonary auscultation analysis is a crucial part of the diagnosis of respiratory illnesses, yet it remains largely a manual task dependent on the skills of individual physicians. Efforts have been made to use machine learning to automate detection of abnormal lung sounds in auscultation analysis, but typical labeled datasets require experts to engage in time-consuming annotation. In contrast, unlabeled data is relatively easy to obtain. In this paper, we use cross pseudo supervision to leverage a small amount of labeled audio data and a larger amount of unlabeled data to perform automated auscultation analysis. We show that this method significantly outperforms fully supervised models trained only on the labeled data.**

***Keywords*—semi-supervised learning, auscultation analysis, pretraining, cross pseudo supervision**

I. Introduction and Problem Definition

Pulmonary auscultation, or the act of listening to lung sounds with a stethoscope, provides vital information to aid in the diagnosis of respiratory diseases by detecting the inhalation-exhalation cycle as well as abnormalities such as crackles, rhonchi, stridor, or wheezes [1]. Despite being non-invasive, it is effective: a review of 28 studies involving a total of 2,032 patients showed that these respiratory sounds are useful indicators of respiratory illnesses such as asthma, cystic fibrosis, and bronchiolitis [2]. Although there have been great advances in healthcare throughout the last century, pulmonary auscultation largely still requires the expertise of medical professionals, due to the high degree of variability present in lung recordings and the difficulty in distinguishing normal from abnormal sounds.

In recent decades, electronic stethoscopes such as the Littmann 3200 have been developed for long-duration auscultation to aid in patient monitoring, but they still require human expertise for analysis; additionally, as physicians may not have the time or auditory acuity to properly analyze the recordings produced, manual analysis can lead to incorrect diagnoses of respiratory diseases and significantly impact patient outcomes [3]. This motivates us to build machine learning models that can automate the process of detecting adventitious sounds in electronic stethoscope recordings as shown in Fig. 1. Note that lung sound labels can overlap, turning this into a multi-label classification problem, which is more difficult than traditional single-label classification tasks.

In addition, traditional machine learning models require large datasets of labeled data. While lung sound recordings are relatively easy to obtain in a hospital setting, labeling the recordings requires the expertise of medical experts and large amounts of annotation time. Therefore, we apply semi-supervised techniques to this problem by training models on a small set of labeled auscultation data and a larger set of unlabeled data, to show the viability of such an approach in lung auscultation analysis.

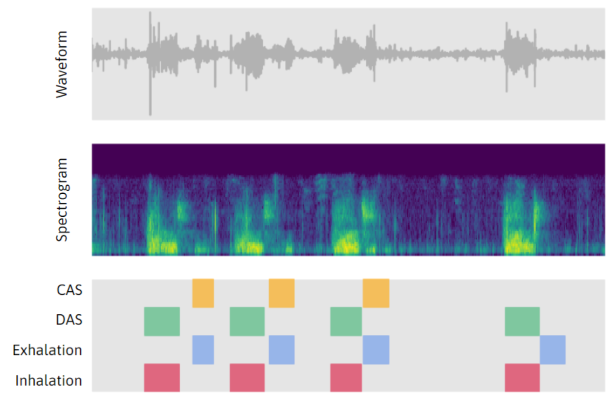

Fig. 1. Illustration of lung sound recording and associated labels. From top to bottom: audio waveform, spectrogram, and annotated labels (CAS: continuous adventitious sound, DAS: discontinuous adventitious sound).

Specifically, we propose a semi-supervised pulmonary auscultation analysis method based on cross pseudo supervision, using two differently initialized networks that supervise each other on unlabeled data. Furthermore, since a model that is trained from scratch will first have to learn to understand audio data before it can learn to predict respiratory sounds, we leverage model backbones pretrained with existing large-scale audio datasets.

Therefore, our contributions are as follows:

- We apply semi-supervised learning techniques using two networks to the task of pulmonary auscultation analysis. To the best of our knowledge this is the first time such techniques have been used for this problem. Particularly, we show that semi-supervised learning with cross pseudo supervision outperforms purely supervised baselines.
- We show that pretraining models on large-scale audio datasets can further improve performance in pulmonary auscultation analysis.

II. Related Work

A. Semi-Supervised Learning

Semi-supervised techniques allow models to leverage a large unlabeled dataset along with a smaller number of labeled examples, which is useful in scenarios such as our problem setting, where obtaining labeled data is expensive but unlabeled data is relatively abundant. One such technique is self-training, where a model predicts pseudo labels for unlabeled data, which is then used to retrain the model. Another technique is label propagation, a graph-based method that assigns labels to unlabeled data on the assumption that similar data has similar labels.

In the last few years, deep learning based approaches have emerged that utilize two networks. Mean Teacher [4] is an approach involving a student network and a teacher network, where the teacher network is the exponential moving average of past student networks. During training, the student network first undergoes supervised training with labeled data, and then is trained with unlabeled data by using pseudo labels generated by the teacher network. Guided Collaborative Training (GCT) [5] is a complex approach that essentially uses two differently-initialized networks and enforces a consistency loss between their predictions, encouraging them to produce similar outputs.

More recently, Cross Pseudo Supervision (CPS) [6] also uses two differently-initialized networks, but instead of directly using network predictions to calculate the consistency loss, it generates pseudo labels from those predictions and calculates consistency loss between each network's prediction and the other network's pseudo labels. Despite its simplicity, the CPS approach has been shown to outperform other semi-supervised methods, including both Mean Teacher and GCT, in semantic segmentation tasks.

B. Automatic Pulmonary Auscultation Analysis and Sound Event Detection

Many machine learning methods have been applied to the task of pulmonary auscultation analysis. Early machine-learning based methods used multilayer perceptrons [7] and convolutional neural networks [8]. Fernando et al. [9] used temporal convolutional networks (TCNs), creating a lightweight yet performant classifier. Hsu et al. [10] used CNNs coupled with LSTMs and GRUs due to the nature of the audio domain, since recurrent networks are effective on sequential data where there are strong relationships between previous and future events.

There exists prior work that uses a similar problem setting to this paper. Chamberlain et al. [11] used a two-stage approach to use labeled data in conjunction with unlabeled data, by building a feature extractor trained on unlabeled data that generates embeddings for a classifier trained on labeled data. Lang et al. [12] use a graph-based method similar to label propagation to leverage unlabeled data. Our approach, using two networks that supervise each other with unlabeled data, enables semi-supervised learning while using labeled and unlabeled data in tandem, rather than in two distinct components.

The Detection and Classification of Audio Scenes and Events (DCASE) Challenge [13] includes a similar task, semi-supervised sound event detection. This is a similar problem, which involves determining the sounds present in a recording and their respective start and end times. One approach uses sound separation techniques that aim to separate individual sound sources from an audio mixture; by isolating specific sound events from the mixture, it becomes easier to detect and classify them [14]. Additionally, some researchers have modified and applied the Detection Transformer (DETR) approach to audio event detection [15]. These are interesting approaches that are outside the scope of this paper, but may be relevant in future research. However, our problem is more difficult because all sounds recorded are human-produced, so the class separation is smaller (i.e., there is a larger disparity between human speech and the hum of a vacuum cleaner than between the sounds of inhalation and exhalation). Therefore, distinguishing respiratory sounds requires a model capable of detecting much more nuanced differences in audio data.

C. Network Pretraining

In recent years, pretraining has proven to be a powerful technique for enhancing the performance of deep neural networks in various domains. The theory is that a network can be first trained on a larger dataset to learn to create useful representations from data, before being fine-tuned for specific tasks. Notably, pretraining on large-scale image datasets such as ImageNet [16] has shown success in improving the performance of image-based machine learning models, as well as accelerating model convergence during training. Models such as ResNet50 [17] pretrained on ImageNet are commonly used as backbones in computer vision models.

Similarly, AudioSet [18] is a large-scale dataset for audio-based applications, enabling the creation of pretrained audio neural networks (PANNs) [19] as feature extractors for the audio domain.

III. Approach

Our model uses a twin network architecture, with each network consisting of a feature extractor generating embeddings from audio spectrogram data, and a multi-label classifier that detects respiratory sound events using those embeddings. The two networks supervise each other on both labeled and unlabeled data; the proposed overall network structure is shown in Fig. 2.

A. Cross Pseudo Supervision

Cross Pseudo Supervision [6] enables semi-supervised learning using two networks that are structurally identical but differently initialized. It enforces consistency regularization by having the networks each generate pseudo labels from unlabeled data, which are then used to supervise the other network. Whereas [6] uses one-hot pseudo labels for semantic segmentation, this is infeasible in our problem setting because multiple classes of sounds may be present at the same time.

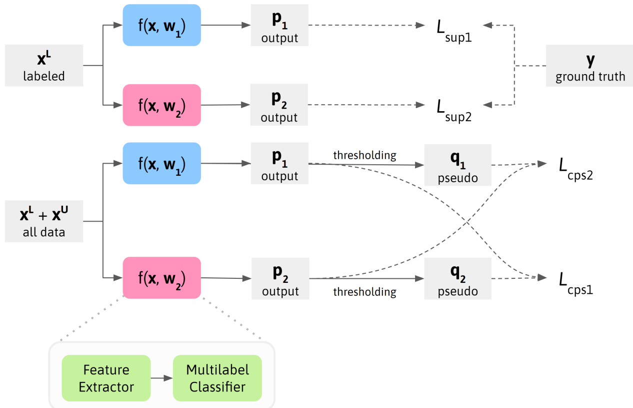

Fig. 2. Overall Model Architecture.

Instead, we generate pseudo labels via thresholding. The CPS loss is then calculated as follows:

$$\mathcal{L}_{cps} = \frac{1}{N}\sum_{n=1}^{N} \ell_{bce}(f(\mathbf{x}_n, w_1), T_\tau(f(\mathbf{x}_n, w_2))) + \ell_{bce}(f(\mathbf{x}_n, w_2), T_\tau(f(\mathbf{x}_n, w_1))), \tag{1}$$

where $f(\mathbf{x}_n, w_k)$ is the output of a network initialized with weights w_k and given input $\mathbf{x}_n$, N is the number of samples (both labeled and unlabeled), $T_\tau(\theta)$ is a thresholding function that returns 1 for $\theta > \tau$ and 0 otherwise, and ℓ_{bce} is the weighted binary cross-entropy loss function, defined as

$$\ell_{bce}(\mathbf{p}, \mathbf{q}) = -\frac{1}{CD}\sum_{c=1}^{C} W_c \sum_{d=1}^{D} \mathbf{q}_{c,d} \cdot \log(\mathbf{p}_{c,d}) + (1 - \mathbf{q}_{c,d}) \cdot \log(1 - \mathbf{p}_{c,d}), \tag{2}$$

where C is the number of output classes, D is the time dimension, W_c is the weight of class c in the loss calculation, $\mathbf{p}$ is the output probabilities, and $\mathbf{q}$ is the labels (which are pseudo labels in the case of CPS loss). We use binary cross-entropy instead of the cross-entropy loss used in [6] due to the multi-label problem setting. To deal with large class imbalances, we weight the binary cross entropy losses for each class using $W_c = F_{max}/F_c$, where F_{max} is the number of time frames in the labeled dataset containing the most common class and F_c is the number of time frames containing class c.

The supervised loss is defined as

$$\mathcal{L}_{sup} = \frac{1}{N_{lab}}\sum_{n=1}^{N_{lab}} [\ell_{bce}(f(\mathbf{x}_n, w_1), \mathbf{y}_n) + \ell_{bce}(f(\mathbf{x}_n, w_2), \mathbf{y}_n)], \tag{3}$$

where N_{lab} is the number of labeled samples in the dataset and $\mathbf{y}$ is the ground truth labels. The final loss is the weighted sum of supervised loss and CPS loss:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{cps}, \tag{4}$$

where λ is a weighting parameter to balance these two losses. The cross supervision using pseudo labels drives the networks towards the same extreme when their outputs are similar, improving the clarity of the decision boundary. If the network outputs are dissimilar, this effect is negated, so unlabeled samples with inconsistent pseudo labels are de-emphasized. As the training process continues and the pseudo labels improve, the unlabeled data effectively helps to expand the training dataset.

B. *Feature Extractor and multi-label Classifier*

Log-mel spectrograms generated from audio data are fed into a feature extractor module, which produces an intermediate representation that is passed to a multi-label classifier module as shown in Fig. 3. We use the CNN architecture from PANN [19] with or without AudioSet pretraining as our feature extractor. Since PANN is designed primarily for clip-level audio source classification and not framewise sound event detection, we modify the pretrained model with deconvolutional layers to increase the model's time resolution. The multi-label classifier module consists of two bidirectional gated recurrent unit (BiGRU) layers coupled to a linear output layer, followed by an interpolator to match the dimensions of the outputs and labels. GRUs are well-suited for time-domain tasks due to their ability to capture long-range temporal dependencies and patterns in sequential data. The bidirectional aspect of a BiGRU allows it to consider past and future information simultaneously, which further improves its ability to find relationships between frames in the time domain.

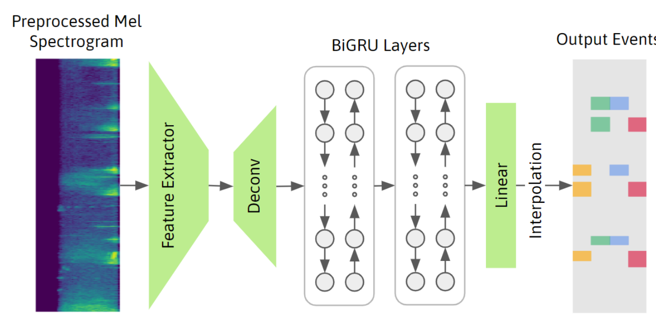

Fig. 3. Single Network Architecture.

IV. Experimental Results

A. *Setup*

Datasets: We train and evaluate our models on *HF_Lung_V1* [10], the largest publicly available dataset of lung sound recordings, as opposed to smaller datasets such as *ICBHI* [20]. It comprises 9,765 strongly labeled 15-second sound recordings from 279 patients. The annotations contain 34,095 inhalation labels, 18,349 exhalation labels, 13,883 continuous adventitious sound labels (comprising 4,740 rhonchi labels, 8,457 wheeze labels, and 686 stridor labels), and 15,606 discontinuous adventitious (crackling) sound labels. The dataset contains recordings from two different machines: a Littmann 3200 electronic stethoscope and an HF-Type-1,

TABLE I

ABLATION STUDY (MODELS TRAINED ON 1/8 OF LABELED DATA; HIGHER NUMBERS ARE BETTER)

Component				AUC				F1			
CNN	**BiGRU**	**Pretrain**	**CPS**	**Inhalation**	**Exhalation**	**DAS**	**CAS**	**Inhalation**	**Exhalation**	**DAS**	**CAS**
✓				89.48%	74.22%	79.30%	84.36%	70.23%	32.61%	35.30%	32.17%
✓	✓			92.64%	79.10%	78.37%	86.49%	73.62%	38.57%	34.98%	34.08%
✓	✓	✓		92.44%	80.77%	83.79%	92.65%	73.43%	40.52%	44.19%	49.69%
✓	✓	✓	✓	**93.59%**	**83.99%**	**87.64%**	**93.88%**	**75.80%**	**45.76%**	**49.16%**	**53.57%**

a recording device custom-built by the dataset authors. We only use recordings from the Littmann as only 18 out of 279 patients were recorded using the HF-Type-1 device.

Evaluation: We evaluate performance using the area under the receiver operating characteristic curve (AUC) and the F1-score. For each model, we calculate AUC and F1 for each of the four classes: inhalation, exhalation, continuous adventitious sound (CAS), and discontinuous adventitious sound (DAS). Following the dataset evaluation setup [10], we group rhonchi, wheeze, and stridor into a single class. We use the train/test split provided by the dataset authors, so there is no cross-contamination between training and testing data (i.e., each patient's data is present only in the training or in the testing set). We use a validation split comprising 10% of the total training data, also taking care to avoid cross-contamination. In total, 2,929 recordings (labeled and unlabeled) are used for training, 325 recordings are used for validation, and 1,250 recordings are used for evaluation.

Data Preprocessing: To reduce noise in the audio data, we first run raw audio waveforms through a tenth-order Butterworth high pass filter to remove frequency data below 100Hz, as that is the lower frequency bound of lung sounds. We then generate log-mel spectrograms from these audio waveforms using 64 Mel bins, a short-time Fourier transform (STFT) window of 256 samples and a hop size of 80 samples. Finally, we conduct min-max normalization on the spectrograms.

Implementation: We implement our models using the PyTorch framework. We initialize the weights of convolutional and linear layers with Kaiming initialization and initialize BiGRUs with orthogonal initialization. For pretrained models, the feature extractor is initialized with the weights of a PANN trained on 8kHz data from AudioSet. We train using the AdamW optimizer with a weight decay of 0.001, a polynomial learning rate (initialized at 0.0001) with power 0.9, and a batch size of 16. We set λ to 1.0 in all experiments. Because Cross Pseudo Supervision generally makes the model training less stable, for CPS models we first start with 50 epochs of non-CPS training, then take the model with the best sum of F1 scores on the validation set to continue with 50 epochs of CPS training using both labeled and unlabeled training data.

B. Results

Ablation Study: We show the results of adding each individual component of our auscultation analysis model in Table I; overall, they each generally improve the model performance. The addition of BiGRU layers produces significant improvement on inhalation and exhalation, demonstrating the BiGRU model's ability to exploit time-wise relationships due to the strong causal correlation between these two classes. Pretraining on AudioSet results in a large improvement in both AUC and F1 on the DAS and CAS classes, showing that this pretraining improves the model generalizability in low-data situations, as these are the two classes with the fewest number of training examples. Cross Pseudo Supervision leads to substantial increases across all classes, showing its ability to leverage the unlabeled data to expand the training dataset via pseudo-labeling.

TABLE II

COMPARISON OF F1 SCORES ACROSS VARIOUS PARTITIONS OF LABELED DATA FOR BASELINE AND PROPOSED MODELS

Partition	Model	Inhalation	Exhalation	DAS	CAS
1/2	Baseline	75.87%	46.59%	46.24%	42.06%
	Proposed	**76.65%**	**49.90%**	**50.46%**	**55.63%**
1/4	Baseline	74.52%	44.19%	41.13%	40.80%
	Proposed	**75.50%**	**47.09%**	**50.55%**	**53.95%**
1/8	Baseline	73.62%	38.57%	34.98%	34.08%
	Proposed	**75.80%**	**45.76%**	**49.16%**	**53.57%**
1/16	Baseline	71.62%	37.80%	30.24%	25.69%
	Proposed	**74.31%**	**43.19%**	**44.67%**	**53.10%**

Comparison with Baselines: We compare our semi-supervised proposed model to the supervised baseline model (which is similar to *CNN+BiGRU*, the best model evaluated in [10]) on $1/2, 1/4, 1/8$, and $1/16$ labeled data partitions using the AUC in Fig. 4 and using the F1-score in Table II. The proposed model outperforms the baseline in all partitions. We observe that the supervised baseline performance drops significantly as the amount of labeled data decreases, while the performance of the proposed model does not degrade as dramatically. On a few occasions, a proposed model trained on lesser amounts of labeled data slightly outperforms a proposed model trained on more labeled data; we hypothesize that this is because the size of the total training dataset (both labeled and unlabeled) is the same when using semi-supervised learning.

Qualitative Analysis: We provide some sample model outputs in Fig. 5 to qualitatively verify the results. We see that the semi-supervised model provides results much closer to the ground truth labels compared to the baseline, especially in classes with fewer examples (namely, DAS and CAS).

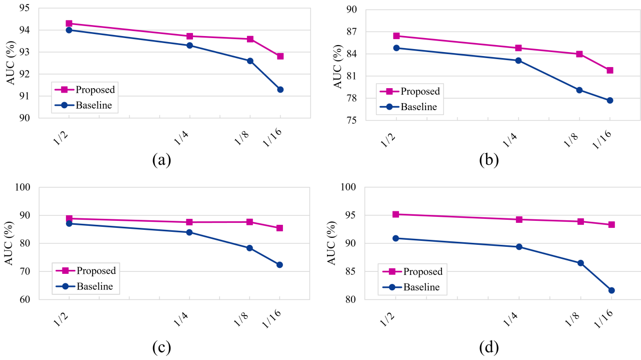

Fig. 4. AUC performance of supervised baselines (blue circles) and the proposed method (pink squares) for all four classes: (a) Inhalation, (b) Exhalation, (c) DAS, (d) CAS.

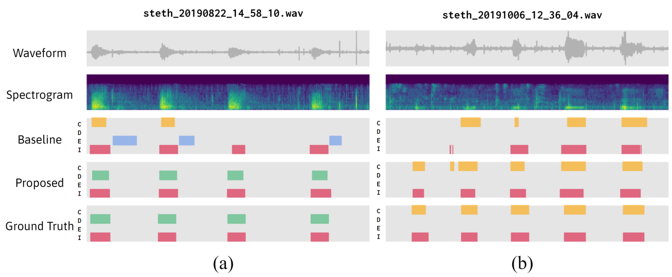

Fig. 5. Qualitative results for baseline and proposed methods trained on (a) 1/2 of labeled data and (b) 1/8 of labeled data.

V. Conclusion

In this work, we have shown that Cross Pseudo Supervision, coupled with network pretraining and BiGRUs to understand temporal relationships, outperforms purely supervised baselines in semi-supervised pulmonary auscultation analysis. The improvements brought on by the usage of semi-supervised techniques show that they hold significant promise in lung auscultation analysis. The dataset used in this study was necessarily fully labeled, in order to test different partitions of labeled and unlabeled data; however, by demonstrating that unlabeled data can be effectively utilized, this study allows for the generation of much larger unlabeled datasets to further enhance the performance of pulmonary auscultation analysis systems. There are avenues for future work; while CPS generally leads to better performance on the evaluation metrics, its addition makes training more unstable since generated pseudo labels may be incorrect, sometimes leading to rapid deterioration of model performance after a period of training. Reweighting approaches such as those proposed in [21] could enable the model to better detect and either disregard or de-emphasize poor pseudo labels. We are confident that continuing research into semi-supervised auscultation analysis will lead to more robust and effective diagnostic tools for respiratory conditions, benefiting both patients and healthcare providers.

References

[1] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, "Auscultation of the respiratory system," Annals of Thoracic Medicine, vol. 10, no. 3, pp. 158–168, 2015.

[2] V. Abreu, A. Oliveira, J. Alberto Duarte, and A. Marques, "Computerized respiratory sounds in paediatrics: A systematic review," Respiratory Medicine: X, vol. 3, p. 100027, Nov. 2021.

[3] J. Heitmann et al., "DeepBreath—automated detection of respiratory pathology from lung auscultation in 572 pediatric outpatients across 5 countries," NPJ Digital Medicine, vol. 6, no. 1, Jun. 2023.

[4] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," Neural Information Processing Systems, 2017.

[5] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. Lau. "Guided collaborative training for pixel-wise semi-supervised learning," in European Conference on Computer Vision (ECCV), 2020.

[6] X. Chen, Y. Yuan, G. Zeng and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 2613-2622, 2021.

[7] S. Rietveld, M. Oud, and E. H. Dooijes, "Classification of asthmatic breath sounds: Preliminary results of the classifying capacity of human examiners versus artificial neural networks," Computers and Biomedical Research, vol. 32, no. 5, pp. 440–448, Oct. 1999.

[8] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," EURASIP Journal on Image and Video Processing, vol. 2017, no. 1, 2017.

[9] T. Fernando, S. Sridharan, S. Denman, H. Ghaemmaghami and C. Fookes, "Robust and interpretable temporal convolution network for event detection in lung sound recordings," in IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 7, pp. 2898-2908, 2022.

[10] F.-S. Hsu 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, p. e0254134, Jul. 2021.

[11] D. Chamberlain, R. Kodgule, D. Ganelin, V. Miglani and R. R. Fletcher, "Application of semi-supervised deep learning to lung sound analysis," 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 2016.

[12] R. Lang, Y. Fan, G. Liu, and G. Liu, "Analysis of unlabeled lung sound samples using semi-supervised convolutional neural networks," Applied Mathematics and Computation, vol. 411, p. 126511, 2021.

[13] N. Turpault, R. Serizel, A. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," In Workshop on Detection and Classification of Acoustic Scenes and Events, New York City, USA, 2019.

[14] I. Kavalerov et al., "Universal sound separation," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 175–179, 2019.

[15] Z. Ye et al., "Sound Event Detection Transformer: An event-based end-to-end model for sound event detection," arXiv.org, Nov. 11, 2021. https://arxiv.org/abs/2110.02011

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[17] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

[18] J. F. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 776-780.

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

[20] B. Rocha et al., "An open access database for the evaluation of respiratory sound classification algorithms," Physiological Measurement, 2019.

[21] Z. Ren, R. Yeh, and A. Schwing, "Not all unlabeled data are equal: Learning to weight data in semi-supervised learning," in Neural Information Processing Systems, 2020.