Pretraining Respiratory Sound Representations using Metadata and Contrastive Learning

Ilyass Moummad, Nicolas Farrugia

PRETRAINING RESPIRATORY SOUND REPRESENTATIONS USING METADATA AND CONTRASTIVE LEARNING

Ilyass Moummad, Nicolas Farrugia

IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France

ABSTRACT

Methods based on supervised learning using annotations in an end-to-end fashion have been the state-of-the-art for classification problems. However, they may be limited in their generalization capability, especially in the low data regime. In this study, we address this issue using supervised contrastive learning combined with available metadata to solve multiple pretext tasks that learn a good representation of data. We apply our approach on respiratory sound classification. This task is suited for this setting as demographic information such as sex and age are correlated with presence of lung diseases, and learning a system that implicitly encode this information may better detect anomalies. Supervised contrastive learning is a paradigm that learns similar representations to samples sharing the same class labels and dissimilar representations to samples with different class labels. The feature extractor learned using this paradigm extract useful features from the data, and we show that it outperforms cross-entropy in classifying respiratory anomalies in two different datasets. We also show that learning representations using only metadata, without class labels, obtains similar performance as using cross entropy with those labels only. In addition, when combining class labels with metadata using multiple supervised contrastive learning, an extension of supervised contrastive learning solving an additional task of grouping patients within the same sex and age group, more informative features are learned. This work suggests the potential of using multiple metadata sources in supervised contrastive settings, in particular in settings with class imbalance and few data.

Index Terms— Audio deep learning, respiratory sound classification, supervised contrastive learning, metadata

1. INTRODUCTION

The main idea of Self-Supervised Learning (SSL) is to solve a pretext task to learn a feature extractor producing useful representations without using labels. Contrastive methods are a family of SSL approaches that optimize an encoder to output similar embeddings for different views of the same data, relying on data augmentations (e.g. SimCLR [1]). In audio, prior work using contrastive approaches have proposed to use synthetic mixing of training examples to generate views [2], as well as sound separation [3], or generating pairs by sampling segments from audio clips [4].

Constrastive learning can also be successfully exploited using a framework called Supervised Contrastive Learning (SCL), which uses classification labels only to group positive pairs in a contrastive setting similar to SimCLR, resulting in state of the art classification on several computer vision benchmarks [5]. SCL has also been applied for environmental sound classification [6]. Instead of using class labels, SCL could be applied on other available information to define pretext tasks. Practical applications usually include metadata, e.g. demographics in medical applications. In this paper, we explore this idea for respiratory sound classification.

Respiratory sound classification is the task of identifying the diagnosis of a breathing cycle whether it's normal or abnormal. Diagnosing a respiratory pathology using ML on audio data would reduce the work overload for physicians and medical experts, and make medical examination less prone to error. ICBHI [7] and SPRSound [8] are public datasets for distinguishing between normal breathing, and respiratory anomalies such as crackle, wheeze, rhonchi and stridor. The two datasets contain recordings of thousands of breathing cycles of varying durations, with an imbalanced class distribution. SPRS recordings were made with an electronic sthetoscope on four different locations, as for ICBHI, recordings were made with four different devices (microphone and electronic stethoscopes) on seven different chest locations. These properties make respiratory sound classification a challenging task. Recent methods based on Deep Learning (DL), especially Convolutional Neural Networks (CNNs), have the ability to learn how to extract and combine relevant representations directly from data. Early DL works on ICBHI, such as LungRN+NL [9] explored data augmentation to address the data class imbalance , as well as attention mechanism in follow-up work LungAttn [10] to improve classification accuracy of respiratory sounds. In addition to data augmentation, RespireNet [11] uses a model pretrained on ImageNet, with a device specific finetuning strategy. A very recent work [12] instead proposed spectrum correction to scale the frequency responses of the recording devices, as well as a co-tuning strategy to learn the relationship between source and target categories to improve transfer learning.

Because of class imbalance and different recording settings, there may be hard samples, and training using the cross-entropy loss may be affected by these samples. In this work, we compare cross-entropy training and supervised contrastive training, as done for environmental sound classification task in SoundCLR [6], and show that the combination of both frameworks can further improve respiratory sound classification scores. In addition, we learn representations using only data augmentation in the SimCLR [1] framework, and using both data augmentation and available metadata in the SCL framework.

Ideas from SCL have already been tested in the context of ICBHI. Song et al. [13] used a simplified SCL framework, by sampling a first batch of examples, and according to their class labels, a second batch is constructed so that each example has a corresponding positive example. For negative samples, a fixed number of samples with different class labels is chosen from all other samples in the two batches. However they did not report results on the official split of ICBHI. In this paper, we adapt the original formulation of SCL of constructing pairs, and outperform cross-entropy (CE) training. We test whether a combination of SCL with CE can fur-

ther boost the classification scores on ICBHI official split. Finally, we show the potential of using available metadata in learning useful representations of respiratory sounds. While using metadata with patient identification with SimCLR has been tested in medical image analysis [14], we propose here to extend the SCL framework to multiple heads to help disentangle subspaces corresponding to different metadata on both ICBHI and SPRSound. Our main contributions are summarized as follows :

1. We show that supervised contrastive learning outperforms the cross-entropy training for respiratory sound classification when finetuning a model pretrained on AudioSet.
2. We combine supervised contrastive learning with cross-entropy and show that it outperforms cross-entropy in correctly classifying anomalies (higher sensitivity).
3. We show that supervised contrastive using metadata with or without class labels learn useful representations, and propose an extension with multiple pretext tasks for a performance boost.

2. METHODS

Let f be a Neural Network (NN) encoder, g a NN classifier, $x_i \in X$ the input breathing cycle, and $y_i \in Y$ the class label, the cross entropy loss (CE) is calculated as follows :

$$\mathcal{L}^{CE} = -\sum_i y_i log(g(f(A(x_i)))) \tag{1}$$

where A is a stochastic augmentation function and $i \in \{1...N\}$ with N being batch size. Here, we train $g \circ f$ to predict the respiratory breathing class for a given respiratory cycle (Fig 1a).

Supervised contrastive learning (SCL) consists of learning a classification task in two steps : first, the feature extractor is trained to pull together in the embedding space samples with the same label, and to push away samples with different label. Second, the linear classifier is trained on the frozen representations learned in the first step (Fig 1b). Formally, the first step consists in adding to the encoder f a shallow NN h called a projector (usually a MLP with one hidden layer) that maps representations to the space where the contrastive loss is applied. In the second step, h is discarded (representations before the non linear projector contains more information [1]), then a classifier g' is trained on the frozen representations (output of f) trained in the first step. The supervised contrastive loss (SCL) is calculated as follows :

$$\mathcal{L}^{SCL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{n \in N(i)} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_n / \tau)} \tag{2}$$

where $i \in I = \{1...2N\}$ the index of an augmented sample within a training batch. $\boldsymbol{z}_i = h(f(A(\boldsymbol{x}_i))) \in \mathbb{R}^{D_P}$ where D_P is the projector's dimension. $P(i) = \{p \in I : y_p = y_i\}$ is the set of indices of all positives in the multiviewed batch distinct from i sharing similar label with i, and $|P(i)|$ is its cardinality, $N(i) = \{n \in I : y_n \neq y_i\}$ is the set of indices of all negatives in the multiviewed batch having dissimilar label with i, the $\cdot$ symbol denotes the dot product, and $\tau \in \mathbb{R}^{+*}$ is a scalar temperature parameter.

We extend the SCL framework to solving multiple pretext tasks using multiple heads sharing the same backbone, we call it

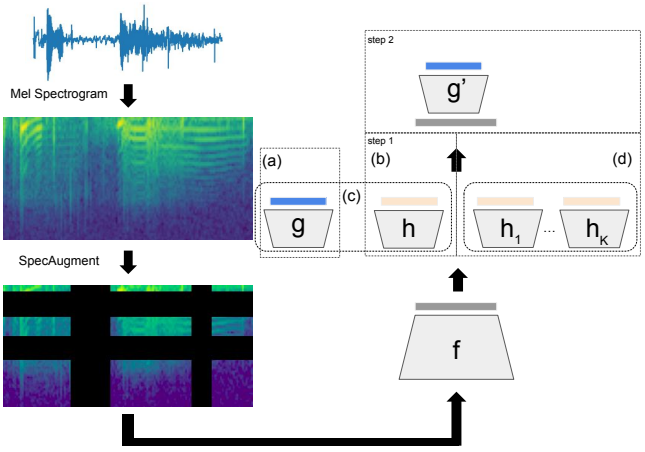

Figure 1: Overview of the proposed framework : cross-entropy training (a), supervised contrastive learning (b), hybrid training (c), multi-supervised contrastive learning (d).

Multi-Supervised Contrastive (M-SCL) (Fig 1d), we define its loss $\mathcal{L}^{M-SCL}$ as :

$$\mathcal{L}^{M-SCL} = \sum_{i \in \{1...K\}} \lambda_i \mathcal{L}_i^{SCL} \tag{3}$$

where K the number of pretext tasks, h_i the projector for the i^{th} task and λ_i the coefficient for the i^{th} loss $\mathcal{L}_i^{SCL}$.

We also consider the hybrid approach, that combines both CE loss and SCL loss (Fig 1c), and minimize the folowing hybrid loss term :

$$\mathcal{L}^{Hybrid} = \alpha \mathcal{L}^{CE} + (1 - \alpha)\mathcal{L}^{SCL} \tag{4}$$

where α controls the tradeoff between the CE and SCL loss terms.

In both M-SCL and Hybrid settings, the backbone f is shared between heads.

3. EXPERIMENTS

3.1. Datasets

ICBHI 2017 [7] is a dataset of the 2017 ICBHI challenge that consists of 5.5 hours of recordings containing 6898 respiratory cycles with a duration ranging from 0.2s to 16.2s (mean cycle duration is 2.7s). 3642 respiratory cycles contain normal breathing, 1864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes, in 920 audio samples from 126 subjects. **SPRSound** [8] is an open-source paediatric respiratory sound database consisting of 2,683 records and 9,089 respiratory sound events from 292 participants. Specifically, there are 6887 normal breathing cycles, 53 rhonchi, 865 wheeze, 17 stridor, 66 coarse crackle, 1167 fine crackle, and 34 of both crackle and wheeze events. There are two tests splits : intra-patient and inter-patient. We consider only the inter-patient test split to test for generalization capabilities of learned representations to unseen patients, a more realistic use case in a clinical setting. These are the two largest respiratory sound datasets, containing additional metadata about patients and acquisitions suited to test our approach of learning representations using metadata.

3.2. Preprocessing and Data Augmentation

The audio recordings are sampled with a rate that varies from 4 kHz to 44.1 kHz for ICBHI and 8 kHz for SPRSound. We re-sample all recordings to 16 kHz mono as done in previous studies [13, 12]. We uniformly limit the maximum duration of a respiratory cycle to 8 seconds. We convert the audio signal into the time-frequency representation Mel-spectrogram, with 64 Mel filterbanks, a window size of 1024 over a hop size of 512, with a minimum and a maximum frequency of 50 and 2000 Hz respectively, because wheezes and crackles are in this interval [15]. We use the data augmentation method SpecAugment [16], that consists of masking blocks of frequency channels and time steps, followed by time warping, that help the network learn features that are robust to partial loss of frequency and time information, and to deformation in the time direction. (Fig 1). SpecAugment has been tested in self-supervised learning of audio representations in previous works [2].

3.3. Model

We use models introduced for the dataset AudioSet [19], a large dataset that contains 2 million sounds including respiratory classes. We compare results with and without pretraining on Audioset. We report results with the CNN6 model, that consists in 4 blocks, each containing a 2D convolution with a kernel size of 5, a batch normalization and an average pooling with a kernel size of 2. CNN6 contains 4.3 million parameters, compared to other works on ICBHI that use ResNet34, ResNet50, and ResNet101 containing approximately 21, 25, and 44 million parameters, respectively[11, 12].

3.4. Evaluation Metrics

We adopt the same evaluation metrics as the official ICBHI 2017 and SPRS challenges :

$$\begin{aligned} Se = \frac{TP}{TP+FN};\ Sp = \frac{TN}{TN+FP};\\ Sc = \frac{Se+Sp}{2};\ HS = 2 * \frac{Se * Sp}{Se+Sp}; \end{aligned} \tag{5}$$

where TP, TN, FP, and FN stand for the numbers of true positives, true negatives, false positives and false negatives, respectively.

3.5. Experimental Setting

We experiment with different setups : CE, SCL, Hybrid, with and without AudioSet pretraining, and M-SCL with 2 pretext tasks with Audioset pretraining, the first pretext task uses respiratory classes, and the second uses synthetic metadata classes (4) : age group (Old/Young for ICBHI and Kid/Baby for SPRSound) and sex (M/F) as these factors can be correlated with presence of respiratory anomalies [20]. We modified the original CNN6 model as follows: we remove dropout, keep all convolutional and pooling layers, then add a simple linear layer for the classifier in the CE setting and the first head of the hybrid setting, and add a MLP with one hidden layer with 128 neurons for the SCL, M-SCL, and the second head of the hybrid setting. We train all models on an Nvidia 3090 GPU for 400 epochs on ICBHI and for 200 epochs on SPRSound (the model converges on it faster) using Adam optimizer with a momentum of 0.9, a batch size of 128 with initial learning rate of 10^{-4} for pretrained models and a learning rate of 10^{-3} for models trained from scratch, and a weight decay of 10^{-4}. We use cosine annealing as a learning rate schedule without warm restarts for all the experiments, except the second phase of supervised contrastive where we train a linear classifier on frozen representations using a learning rate of 0.1 without scheduling. We performed a grid search to find the best parameters for SpecAugment on ICBHI. The best and most stable results were obtained without time warping, when masking two blocks of twenty consecutive mel frequency channels, and two blocks of forty consecutive time steps. We also performed a grid search to find the best value of $\tau = 0.06$ in SCL as well as $\alpha = 0.5$ for hybrid training, and $\lambda_1 = 0.25$ and $\lambda_2 = 0.75$ for M-SCL. We use the same values of hyperparameters for SPRSound. We report results on the ICBHI official challenge split and the SPRSound inter-patient set (as the intra-patient set doesn't reflect generalization capabilities). In order to be comparable with previous approaches, and in the absence of validation split, results correspond to the epoch with the best Sc on the test split as done by previous works [11]. For stability and robustness, we report results over ten identical runs.

3.6. Results

Table 1 is composed of 2 panels : the upper panel shows state-of-the-art performance on the ICBHI official split, and the bottom panel shows our results across 10 identical runs using cross-entropy training, supervised contrastive training with class labels only, and hybrid training for CNN6 both from scratch and with AudioSet pretraining. Our results show that supervised contrastive learning or hybrid learning surpass cross-entropy training when training from scratch or by finetuning from AudioSet. A fine-tuned CNN6 reaches a score of 57.55 in the supervised contrastive setting outperforming all previous work except the setting of the recent work of Nguyen et al. [12], in which they introduced co-tuning technique and obtains a score of 58.29, while our approach compares well with their vanilla finetuning and StochNorm scores. In terms of memory requirement, our approach has up to five times less parameters than previously proposed approaches with networks finetuned from Imagenet. Finally, it is worth noting that our approach with CNN6 trained from scratch with the hybrid loss achieves the new state of the art on ICBHI when not considering pretraining on an external dataset.

For learning representations using metadata, we consider the pretext task of pulling together respiratory cycles sharing the same demographics (sex, and/or age group) while pushing away dissimilar ones in the latent space. We also test a combination of metadata with class labels, using either a single head with a label combining target classes and metadata using SCL or multiple heads (M-SCL). In a second step, we train a linear classifier on the frozen representation using class labels. Table 2 shows the potential of learning representations of data with or without the use of class labels; the first panel shows that using sex and age group with SCL learns representations that obtain a score of 55.29$\pm$0.80, outperforming the SSL baseline SimCLR (48.34$\pm$0.87), and doing as well as cross-entropy training with class labels (Table 1); in the second panel, we can see that M-SCL (58.04$\pm$0.94) outperforms SCL (55.81$\pm$0.88), we note here that both methods use metadata and class labels : M-SCL uses a first head head for class labels and a second head for metadata, while SCL uses one head for the label obtained by combining metadata and classlabels. SCL with label and metadata performs worse than SCL with class labels from Table 1, we hypothesize that it's because for some synthetic classes only few samples are available (e.g. only 3 respiratory cycles with crackles and wheezes come from young female), therefore it's hard to learn discriminative feature for those classes. However, M-SCL outperforms our best result in Ta-

Table 1: Performance analysis on ICBHI.

Method		Sp	Se	Sc	# of Parameters (in M)	Ext. Dataset
LungRN+NL[9]		63.2	**41.3**	52.3	-	None
LungAttn[10]		**71.44**	36.36	**53.9**	0.7	None
Wang et al.[17]		70.4	40.2	55.3	25 (estimated)	ImageNet
RespireNet[11]		72.3	40.1	56.2	21 (estimated)	ImageNet
ARSC-Net[18]		67.13	**46.38**	56.76	-	ImageNet
Nguyen et al.[12] (Vanilla)		76.33	37.37	56.85	23 (estimated)	ImageNet
Nguyen et al.[12] (StochNorm)		78.86	36.40	57.63	23 (estimated)	ImageNet
Nguyen et al.[12] (CoTuning)		**79.34**	37.24	**58.29**	23 (estimated)	ImageNet
Backbone	Method	Our Results (10 runs)				
3*CNN6	CE	**76.72±3.97**	31.12±3.72	53.92±0.71	4.3	None
	SCL	76.17±3.84	27.97±3.92	52.08±1.06	4.3	None
	Hybrid	75.35±5.47	**33.84±5.67**	**54.74±0.5**	4.3	None
3*CNN6	CE	70.09±3.08	40.39±2.97	55.24±0.43	4.3	AudioSet
	SCL	**75.95±2.31**	39.15±1.89	**57.55±0.81**	4.3	AudioSet
	Hybrid	70.47±2.07	**43.29±1.83**	56.89±0.55	4.3	AudioSet

*We highlights in bold our best scores as well as best scores from the litterature both from scratch and from pretraining.

Table 2: Results on IBCHI using metadata

Method	Sp	Se	Sc
without respiratory classes			
SimCLR	59.75±5.81	36.93±5.71	48.34±0.87
SCL w/ Sex+Age	71.25±2.79)	39.32±3.35	55.29±0.80
with respiratory classes			
SCL w/ Age+Class	70.84±3.19	**40.47±3.84**	55.65±1.27
SCL w/ Sex+Class	71.00±4.52	39.72±3.13	55.36±1.00
SCL w/ Sex+Age+Class	71.56±4.03	40.06±3.09	55.81±0.88
M-SCL w/ Sex+Age & Class	**76.93±2.99**	39.15±2.84	**58.04±0.94**

Table 3: Results on SPRSound

Method	Se	Sp	Sc	HS
CE	76.89±0.80	92.35±1.10	84.62±0.29	83.90±0.25
SCL	80.69±1.62	90.49±1.27	85.59±0.48	85.29±0.56
M-SCL	82.24±2.24	88.62±1.48	85.43±0.79	85.27±0.88
Baseline	51.93	77.88	64.90	62.31

ble 1 (SCL : 57.55±0.81), showing the performance boost obtained from leveraging metadata.

Table 3 shows the performance of CNN6 (pretrained on AudioSet) when finetuned on SPRSound using CE, SCL or M-SCL, compared to the baseline of Naive Bayes Classifier trained on the mel-frequency cepstral coefficient (MFCC). Our experiments show that SCL and M-SCL outperform CE in both harmonic score (HS) and mean score (Sc), only the specificity (Sp) is higher for CE, we assume this is due to overfitting the normal breathing cycle because of the imbalance in the dataset. On the contrary, our contrastive approach learns to better separate anomalies in the latent space as the sensitivity (Se) is higher than of CE, especially for M-SCL when leveraging metadata information for learning representations. We did not compare ours results on SPRS to other works other than the baseline, because the few published works report their results by combining both inter and intra sets.

On both datasets, contrastive learning has led to higher score than cross entropy training with high sensitivity. Contrastive learning learns to cluster similar breathings in the latent space while pushing apart dissimilar ones, a desirable property for classification. Metadata provided additional information to be taken into account when learning the representations of breathings. We experimented with sex and age (4 groups in total), and they either boosted sensitivity (on SPRSound) or overall score (on ICBHI). Our approach is distinct from previous work that have used contrastive approaches for respiratory sounds; the work of [20] trained a system on ICBHI to diagnose patients, which is an easier task. [13] have used SCL on ICBHI with different sampling of negatives examples, as well as a different cross-validation split, making it difficult to compare. Overall, we have proposed a supervised contrastive approach that exploits metadata in a simple, effective and reproducible way on two datasets.

4. CONCLUSION

We show in this work the potential of supervised contrastive learning for an imbalanced and noisy setting, outperforming cross-entropy using experiments on respiratory sound classification. We also show that using metadata to combine multiple supervised contrastive tasks for learning useful representations obtain state-of-the-art results. In future work, we will attempt at building upon the multi-head framework using several pretext tasks such as exploiting spatial or temporal metadata associated with recordings, and investigate data augmentation techniques that better address the variability in the low data regime. Ultimately, such approaches could be adapted to generalize to unseen auditory tasks such as detection or localization, and deal with larger domain shifts.

Acknowledgment

This work was co-funded by the AI@IMT program of the ANR (French National Research Agency) and the company OSO-AI. We would like to thank our colleagues in the BRAIn Team of the Mathematical and Electrical Engineering Department of Institut Mines-Télécom Atlantique for their insights and feedback.

5. REFERENCES

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, 2020.

[2] E. Fonseca, D. Ortego, K. McGuinness, N. E. O'Connor, and X. Serra, "Unsupervised contrastive learning of sound event

representations," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[3] E. Fonseca, A. Jansen, D. P. Ellis, S. Wisdom, M. Tagliasacchi, J. R. Hershey, M. Plakal, S. Hershey, R. C. Moore, and X. Serra, "Self-supervised learning from automatically separated sound scenes," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[4] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3875–3879.

[5] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, 2020.

[6] A. Nasiri and J. Hu, "Soundclr: Contrastive learning of representations for improved environmental sound classification," 2021, arXiv:2103.01929.

[7] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni, *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, 2019.

[8] Q. Zhang, J. Zhang, J. Yuan, H. Huang, Y. Zhang, B. Zhang, G. Lv, S. Lin, N. Wang, X. Liu, M. Tang, Y. Wang, H. Ma, L. Liu, S. Yuan, H. Zhou, J. Zhao, Y. Li, Y. Yin, L. Zhao, G. Wang, and Y. Lian, "Sprsound: Open-source sjtu paediatric respiratory sound database," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 5, pp. 867–881, 2022.

[9] Y. Ma, X. Xu, and Y. Li, "Lungrn+nl: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation," in *Proc. Interspeech 2020*, 2020.

[10] J. Li, J. Yuan, H. Wang, S. Liu, Q. Guo, Y. Ma, Y. Li, L. Zhao, and G. Wang, "Lungattn: advanced lung sound classification using attention mechanism with dual tqwt and triple stft spectrogram," *Physiological Measurement*, 2021. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/ac27b9

[11] S. Gairola, F. Tom, N. Kwatra, and M. Jain, "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," 2020.

[12] T. Nguyen and F. Pernkopf, "Lung sound classification using co-tuning and stochastic normalization," *IEEE Transactions on Biomedical Engineering*, 2022.

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

[14] Y. N. T. Vu, R. Wang, N. Balachandar, C. Liu, A. Y. Ng, and P. Rajpurkar, "Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation," in *Proceedings of the 6th Machine Learning for Healthcare Conference*, 2021. [Online]. Available: https://proceedings.mlr.press/v149/vu21a.html

[15] N. Jakovljević and T. Lončar-Turukalo, "Hidden markov model based respiratory sound classification," in *Precision Medicine Powered by pHealth and Connected Health*, 2018.

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

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

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

[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*, 2020.

[20] P. N. Soni, S. Shi, P. R. Sriram, A. Y. Ng, and P. Rajpurkar, "Contrastive learning of heart and lung sounds for label-efficient diagnosis," 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389921002671