This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2022.3156293, IEEE Transactions on Biomedical Engineering

Lung Sound Classification Using Co-tuning and Stochastic Normalization

Truc Nguyen, and Franz Pernkopf, *Senior Member, IEEE*

Abstract—**Computational methods for lung sound analysis are beneficial for computer-aided diagnosis support, storage and monitoring in critical care. In this paper, we use pre-trained ResNet models as backbone architectures for classification of adventitious lung sounds and respiratory diseases. The learned representation of the pre-trained model is transferred by using vanilla fine-tuning, co-tuning, stochastic normalization and the combination of the co-tuning and stochastic normalization techniques. Furthermore, data augmentation in both time domain and time-frequency domain is used to account for the class imbalance of the ICBHI and our multi-channel lung sound dataset. Additionally, we introduce spectrum correction to account for the variations of the recording device properties on the ICBHI dataset. Empirically, our proposed systems mostly outperform all state-of-the-art lung sound classification systems for the adventitious lung sounds and respiratory diseases of both datasets.**

Index Terms—**Adventitious lung sound classification, respiratory disease classification, crackles, wheezes, co-tuning for transfer learning, stochastic normalization, ICBHI dataset.**

I. Introduction

RESPIRATORY diseases have become one of the main causes of death in society. According to the World Health Organization (WHO), the ”big five” respiratory diseases, which include asthma, chronic obstructive pulmonary disease (COPD), acute lower respiratory tract infections, lung cancer and tuberculosis, cause the mortality of more than 3 million people each year worldwide. Currently, CoViD-19, a special form of viral pneumonia related to the coronavirus identified firstly in Wuhan (China) in 2019, has caused globally more than 158 million infections and 3,296,000 deaths [1]. On March 11, 2020, the WHO officially announced that CoViD-19 has reached global pandemic status. Furthermore, according to [2], the “big five” lung diseases, except lung cancer, have increased during CoViD-19 epidemics. These respiratory diseases are characterised by highly similar symptoms, i.e. the adventitious breathing, which could be a confounding factor during diagnosis [3]. Due to their severe consequences, an early and accurate diagnosis of these types of diseases has become crucial.

Lung sounds convey relevant information related to pulmonary disorders with adventitious breathing sounds such as crackles and/or wheezes [4], [5]. In the last decades, computational lung sound analysis (CLSA) [6] have been developed to facilitate a more objective assessment of the lung sound for diagnosis of pulmonary diseases/conditions. CLSA systems automatically detect and classify adventitious lung sounds by using digital recording devices, signal processing techniques and machine learning algorithms. They are also carefully evaluated in real-life scenarios and can be used as portable easy-to-use devices without the necessity of expert interaction. Recently, automatic diagnostic of CoViD-19 disease has been popular using respiratory sound data including cough, voice and breaths [7], [8]. Most of the CoViD-19 diagnostic systems use respiratory sound datasets such as Coswara [7], CoViD-19 crowd-sourced sound dataset [9] or COUGHVID [10]. There is only a modest number of works using lung sounds recorded by digital stethoscope. For instance, in [11], the auscultation recordings have been analysed by six physicians but no machine learning has been applied. In [12], an automated lung sound analysis – the LungPass platform – has been introduced. It is based on neural networks for identifying lower respiratory tract involvement in COVID-19. In our paper, we focus on CLSA using only lung sounds from our multi-channel lung sound dataset and the ICBHI 2017 dataset, which is a popular and public lung sound dataset for benchmarking.

In CLSA, there are two popular classification tasks, namely (i) adventitious lung sound and (ii) respiratory disease classification. We consider both in this article. In adventitious lung sound classification, recognition of normal and abnormal sounds (i.e. crackles and/or wheezes) is important; while for respiratory disease classification, several categories have been considered e.g. binary classification (health and pathological), ternary chronic classification (healthy, chronic and non-chronic diseases) or six class classification of distinct pathologies. The proposed systems have been evaluated on non-public datasets such as R.A.L.E. [13] or multi channel lung sound data [14] (ours) and public datasets i.e. the ICBHI 2017 dataset [5] or the Abdullah University Hospital 2020 dataset [15]. Due to limitations in the amount and quality of available data, the performance and generalization of the lung sound classification system may suffer. To deal with these challenges, different feature extraction methods [16], [17], conventional machine learning [18], [19], [20], deep learning [21], [22], [23], [24] and data augmentation have been introduced in the recent past.

Deep neural networks (DNNs) trained from scratch require large amounts of data. As data collecting is a time consuming task for lung sound data, transferring pre-trained parameters from DNNs, which are trained on other datasets e.g. ImageNet is advantageous. Less data of the target task is required, faster training is enabled, and usually better performance after fine-tuning the model on the target task is achieved [25]. Therefore,

This research was supported by the Vietnamese - Austrian Government scholarship.

T. Nguyen is with Signal Processing and Speech Communication Lab. Graz University of Technology, Austria (e-mail: t.k.nguyen@tugraz.at).

F. Pernkopf is with Signal Processing and Speech Communication Lab. Graz University of Technology, Austria (e-mail: pernkopf@tugraz.at).

fine-tuning brings great benefit to the research community such as transfer learning from ImageNet [26], [27], or audio scene datasets [28].

In this work, we further improve the generalization ability and model performance for adventitious lung sound classification and respiratory disease classification systems using the ICBHI 2017 dataset and our multi-channel lung sound dataset. The main contribution is to exploit different transfer learning approaches, in which the pre-trained ResNet models of the ImageNet classification task are used as backbone architectures. We compare the following approaches:

- We fine-tune the pre-trained model on a target domain and update all top (i.e. feature representation) layers and bottom (i.e. task-specific) layers. We call this vanilla fine-tuning.
- We apply *co-tuning* to fully transfer the knowledge of the pre-trained model [29], in which representation layers and task-specific layers of both source domain and target domain are collaboratively exploited. Co-tuning learns a relationship between source and target categories. Both, the target labels and the probabilistic source labels determined by the category relationship are used for fine-tuning the model for the target domain [29].
- We replace Batch Normalization (BN) layers, which suffer from poor performance in case of a data distribution shift between training and test data. We introduce stochastic normalization (StochNorm) [30] in each residual block of the pre-trained backbone architecture. StochNorm is a parallel structure normalizing the activation of each channel by either mini-batch statistics or moving statistics to avoid influence of sample statistics during training. Thus, it is considered as a regularization method. Furthermore, fine-tuning inherits further prior knowledge of moving statistics of the pre-trained networks compared to vanilla fine-tuning. Both properties help to avoid over-fitting on small datasets such as the ICBHI and our lung sound dataset.
- We combine co-tuning and stochastic normalization techniques to take advantages of both techniques.

Furthermore, we apply data augmentation in both time domain and time-frequency domain to account for the class imbalance in the datasets. In particular, beside using time stretching on audio signals, we double the size of the training dataset by flipping samples i.e spectrograms in the target domain. This enhances the performance of adventitious lung sound classification. In addition, we use spectrum correction [31] of the lung sounds to compensate the recording device variations in the ICBHI dataset. This improves the generalization ability by accounting for the recording device differences. Currently, there are a few approaches which address this problem. They focus on either training or fine-tuning specific models for a specific device which is used for majority recordings to limit sensitivity to characteristics of the recording device [32], [33].

The outline of the paper is as follows: In Section II, we introduce the lung sound databases. In Section III, we present our lung sound classification systems. In Section IV, we present the experimental setup including the evaluation metrics and the experimental results. We review related works for ICBHI and our multi-chanel lung sound dataset in Section V. Finally, we conclude the paper in Section VI.

II. Databases

We evaluate our models on the ICBHI dataset and our multi-channel lung sound dataset. Both are introduced in the following.

A. ICBHI 2017 Dataset

The ICBHI 2017 database [5] consists of 920 annotated audio samples from 126 subjects corresponding to patient pathological conditions i.e. healthy and seven distinct disease categories (Pneumonia, Bronchiectasis, COPD, upper respiratory tract infection (URTI), lower respiratory tract infection (LRTI), Bronchiolitis, Asthma). The audios were recorded using different stethoscopes i.e. AKGC417L, Meditron, Litt3200 and LittC2SE. The recording duration ranges from 10s to 90s and the sampling rate ranges from 4000Hz to 44100Hz. Each recording is composed of a certain number of breathing cycles with corresponding annotations of the beginning and the end, and the presence/absence of crackles and/or wheezes. The annotations of the database support to split audio recordings into respiratory cycles. The cycle duration ranges from 0.2s to 16s and the average cycle duration is 2.7s. The database includes 6898 different respiratory cycles with 3642 normal cycles, 1864 crackles, 886 wheezes, and 506 cycles containing of both crackles and wheezes.

We propose a classification system for the following tasks.

- ALSC: Adventitious lung sound classification (ALSC) is separated into two sub-tasks for respiratory cycles. The first one is a 4-class task classifying respiratory cycles into four classes (*Normal*, *Crackles*, *Wheezes* and *both Crackles and Wheezes*). The second sub-tasks is a 2-class task of normal and abnormal lung sounds including *Crackles*, *Wheezes* and *both Crackles and Wheezes*. We evaluate our system on the official ICBHI data split. The dataset was divided by the ICBHI challenge into 60% for training and 40% for testing. Both sets are composed with different patients (i.e. non-overlapping).
- RDC: Respiratory disease classification (RDC) also consists of two sub-tasks for audio recordings. The first one is a 3-class task classifying audio recordings into three groups of *Healthy*, *Chronic Diseases* (i.e. COPD, Bronchiectasis and Asthma) and *Non-Chronic Diseases* (i.e. URTI, LRTI, Pneumonia and Bronchiolitis). The second sub-tasks is a 2-class task (healthy/unhealthy), where the unhealthy class comprises of the seven diseases. Similarly, we evaluate our system on the official ICBHI data split.

Beside the target domain role for ALSC and RDC, the ICBHI dataset is used as an additional source domain to retrain pre-trained models of ImageNet for crackle detection on our multi-channel lung sound dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2022.3156293, IEEE Transactions on Biomedical Engineering

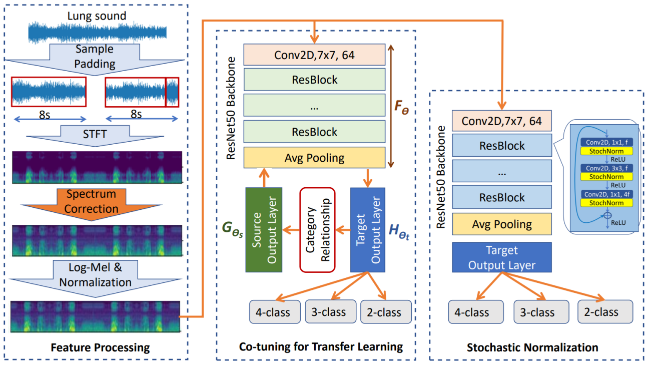

Fig. 1. Proposed systems using co-tuning for transfer learning or stochastic normalization.

B. Multi-channel Lung Sound Database

The multi-channel lung sound database [21], [14] has been recorded in a clinical trial. It contains lung sounds of 16 healthy subjects and 7 patients diagnosed with idiopathic pulmonary fibrosis (IPF). We used our 16-channel lung sound recording device to record lung sounds over the posterior chest at two different airflow rates, with 3 - 8 respiratory cycles within 30s. The lung sounds were recorded with a sampling frequency of 16kHz. The sensor signals are filtered with a Bessel high-pass filter with a cut-off frequency of 80Hz and a slope of 24dB/oct. We extracted full respiratory cycles using the airflow signal from all recordings. We manually annotated respiratory cycles in cooperation with respiratory experts from Medical University of Graz, Austria. The number of normal breathing and crackle cycles of 16 healthy and 7 IPF subjects are 4405 and 1791, respectively. We use this dataset as another target domain to evaluate our proposed systems using transfer learning from different source domains.

III. Proposed Lung Sound Classification Systems

The proposed systems include two key stages i.e. feature processing and classification as shown in Fig. 1. Firstly, the respiratory cycles/ recordings are pre-processed in time domain and transformed into log-mel spectrograms of fixed size. Secondly, the features are fed to the CNN model where co-tuning or stochastic normalization are explored for the different classification tasks. During inference, the label of an input respiratory cycle/ recording is determined via majority voting [34] of the predicted labels of the individual segments.

A. Audio Pre-processing and Feature Extraction

We use the audio pre-processing and feature extraction techniques presented in [24] for both datasets. Audio recordings are resampled to 16kHz for the ALSC tasks of the ICBHI challenge and our dataset, while the RDC tasks use 4kHz sampling rate. Similar to our previous works on ALSC of ICBHI and our multi-channel dataset [24], [35], the respiratory cycles are split without overlap into segments. Furthermore, we apply sample padding in time-reversed order to achieve fixed-length segments without abrupt signal changes. For the RDC task of the ICBHI dataset, recordings are decomposed into segments of the same length using 50% overlap. An ablation study over different segment lengths is provided in Section IV. Again sample padding is applied to the segments being shorter than the fixed length. Hence, the pre-processing for both tasks is similar.

We use a window size of 512 samples for the fast Fourier transform (FFT) using 50% overlap between the windows. The number of mel frequency bins is chosen as 50 and 45 for the ICBHI dataset and our multi-channel dataset, respectively. The logarithmic scale is applied to the magnitude of the mel spectrograms. The log-mel spectrograms are normalized with zero mean and unit variance. Then these spectrograms are duplicated into three channels to match the input size of the pre-trained ResNet model for the ALSC task. However, for the RDC task of the ICBHI dataset, we convert the spectrogram, which is considered as a grey image, into a RGB color image and enlarge the image to twice the size using linear interpolation. These techniques are commonly used [36].

B. Spectrum Correction

We observe a different frequency response across recording stethoscopes which results in a performance degradation for under-represented devices. Hence, we calibrate the features of the audio segments by applying spectrum correction instead of training or fine-tuning the model for a specific device [32], [33]. The spectrum correction or calibration was first applied for acoustic scene classification [31]. It scales the frequency response of the recording devices. In particular, the calibration coefficients are calculated for each device based on data from reference devices. The recorded data portions of recording devices i.e. *AKGC417L*, *Meditron*, *Litt3200* and *LittC2SE* of the ICBHI dataset are 63%, 21%, 9% and 7%, respectively. The magnitude spectrum $\boldsymbol{s}_i^k$ of each segment i recorded by the device k is an averaged spectrum along the time axis of all FFT windows. The mean device spectrum $\bar{\boldsymbol{s}}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} \boldsymbol{s}_i^k$, where device k records N_k segments corresponding to N_k spectra. The reference spectrum $\boldsymbol{s}_{ref}$ is furthermore averaged over all mean device spectra of the D reference devices $\boldsymbol{s}_{ref} = \frac{1}{|D|}\sum_{k \in D} \bar{\boldsymbol{s}}_k$, where D contains the indices of the reference devices. We investigate different cases of reference devices based on their prominence i.e only one device either *AKGC417L* or *Meditron* or both *AKGC417L* and *Meditron*, or all recording devices. The scaling coefficients $\boldsymbol{c}_k$ of each device is the element-wise fraction (i.e for each frequency bin) of the reference spectrum and its corresponding device spectrum $\boldsymbol{c}_k = \frac{\boldsymbol{s}_{ref}}{\bar{\boldsymbol{s}}_k}$. The magnitude of the STFTs of each device is scaled by using the corresponding coefficient vector $\boldsymbol{c}_k$ for the frequency bins. We empirically observed that the normalization in spectrogram domain is more successful than in log-mel domain.

C. Data Augmentation

The ICBHI 2017 dataset is extremely imbalanced with around 53% of respiratory cycles belonging to the normal class and 86% of audio recordings belonging to COPD. Furthermore, with our multi-channel lung sound dataset, around 71%

of respiratory cycles are annotated as normal class. Therefore, we use data augmentation in both time domain and time - frequency domain in order to balance the training dataset and prevent over-fitting.

1) Time Domain: For ALSC of the ICBHI dataset, we use time stretching to increase/reduce the sampling rate of an audio signal without affecting its pitch [37]. It is used to double the number of segments of the wheeze, and both wheeze and crackle classes. We use a random sampling rate uniformly distributed with ±10% of the original sampling rate. For RDC of ICBHI, time stretching is used for all classes to double the number of samples. Furthermore, on the doubled training set further data augmentation methods[1] i.e volume adjusting, noise addition, pitch adjusting and speed adjusting are randomly applied based on a predefined probability.

2) Time-Frequency Domain: Vocal tract length perturbation (VTLP) selects a random wrap factor α for each recording and maps the frequency f of the signal bandwidth to a new frequency f' [38]. We select α from a uniform distribution $\alpha \sim \mathcal{U}(0.9, 1.1)$ and set the maximum signal bandwidth to F_{hi} = [3200, 3800]. VTLP is applied directly to the mel filter bank rather than distorting each spectrogram. VTLP is applied to enlarge the dataset for all classes in both tasks for both the original training set and the time stretched data. Additionally, we double the log-mel features by adding the flipped log-mel features (in frequency axis) for the ALSC and crackle detection task of our dataset.

D. Exploiting Transferred Knowledge

1) Transfer Learning: Given a DNN M_0 pre-trained on a source dataset $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{m_s}$, transfer learning aims to fine-tune M_0 on a target dataset $\mathcal{D}_t = \{(x_t^i, y_t^i)\}_{i=1}^{m_t}$. In this work, $\mathcal{D}_s$ is selected from ImageNet and $\mathcal{D}_t$ is the ICBHI 2017 dataset or our multi-channel lung sound dataset. Only $\mathcal{D}_t$ and the pre-trained model M_0 are available during fine-tuning. Because $\mathcal{D}_s$ and $\mathcal{D}_t$ are different domains, which may have different input spaces $\mathcal{X}_s$ and $\mathcal{X}_t$, corresponding to different output spaces $\mathcal{Y}_s$ and $\mathcal{Y}_t$, respectively. Therefore, M_0 can not be directly applied to the target data. It is common practice, to split M_0 into two parts: a general representation function $F_{\bar{\theta}}$ (parametrized by $\bar{\theta}$) and a task-specific function G_{θ_s} (parameterized by θ_s), which denotes the last layers of the pre-trained model. Usually, the representation function is retained and the task-specific function is replaced by a randomly initialized function H_{θ_t} (parameterized by θ_t) whose output space matches $\mathcal{Y}_t$. Hence, we optimize

$$(\bar{\theta}^*, \theta_t^*) = \underset{\bar{\theta}, \theta_t}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_t|} \sum_{i=1}^{m_t} l(H_{\theta_t}(F_{\bar{\theta}}(x_t^i)), y_t^i), \tag{1}$$

where $l(\cdot)$ is a loss function such as cross-entropy for classification. We will call this vanilla fine-tuning. Pre-trained parameters $\bar{\theta}$ provide a good starting point for the optimization. It means that the vanilla fine-tuning for a target dataset can be beneficial by transferring the knowledge of the part $F_{\bar{\theta}}$ of the source dataset.

In this work, we explore different depths of ResNet architectures i.e. ResNet18, ResNet34, ResNet50 and ResNet101 as neural network backbones.

2) Co-tuning: Co-tuning for transfer learning enables full knowledge transfer of the pre-trained models using a two-step framework [29]. The first step is learning the relationship between source categories and target categories from the pre-trained model with calibrated predictions. Secondly, target labels (one-hot labels) and source labels (probabilistic labels) translated by the category relationship, collaboratively supervise the fine-tuning process. Co-tuning empirically proves its ability in enhancement of the performance compared to vanilla fine-tuning of the ImageNet pre-trained models [29]. In this work, we apply co-tuning to fully exploit the ImageNet pre-trained models for significantly distinct datasets such as the ICBHI and our multi-channel lung sound dataset. The co-tuning block in Fig. 1 shows the source output layer G_{θ_s}, the target output layer H_{θ_t}, the ResNet50 backbone $F_{\bar{\theta}}$ and category relationship, which is the relationship between output spaces i.e. the conditional distribution $p(y_s|y_t)$.

During training, the category relationship $p(y_s|y_t)$ is needed to translate target labels y_t into probabilistic source categories y_s, which is used to fine-tune the task-specific function G_{θ_s}. The gradient of G_{θ_s} can be back-propagated into $F_{\bar{\theta}}$. Both outputs y_t and y_s collaboratively supervise the transfer learning process described as

$$\begin{aligned}(\bar{\theta}^*, \theta_t^*, \theta_s^*) = \underset{\bar{\theta}, \theta_t, \theta_s}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_t|} \sum_{i=1}^{m_t} [l(H_{\theta_t}(F_{\bar{\theta}}(x_t^i)), y_t^i) \\ + \lambda l(G_{\theta_s}(F_{\bar{\theta}}(x_t^i)), p(y_s|y_t = y_t^i))],\end{aligned} \tag{2}$$

where λ trades off the target and source supervisions. Variables $\bar{\theta}$ and θ_s are initialized from pre-trained weights. In this way, the pre-trained parameters are fully exploited in the collaborative training. During inference, the task specific layers G_{θ_s} are removed to avoid the additional cost.

The category relationship $p(y_s|y_t)$ is computed based on the output of task-specific function G_{θ_s} (i.e. a probability distribution over source categories $\mathcal{Y}_s$) and target labels $\mathcal{Y}_t$ by two ways:

- Direct approach: The category relationship is determined as average of the predictions of the pre-trained source model over all samples of each target category i.e.

$$p(y_s|y_t = y) \approx \frac{1}{|\mathcal{D}_t^y|} \sum_{(x, y_t) \in \mathcal{D}_t^y} M_0(x), \tag{3}$$

  where $\mathcal{D}_t^y = \{(x, y_t) \in \mathcal{D}_t | y_t = y\}$ and the pre-trained model M_0 is considered as a probabilistic model approximating the conditional distribution $M_0(x) \approx p(y_s|x)$.
- Reverse approach: When categories in the pre-trained dataset are diverse enough to compose a target category, we can use a reverse approach. We learn the mapping $y_s \to y_t$ from $(M_0(x_t), y_t)$ pairs, where y_t is the target label and $M_0(x) \approx p(y_s|x)$ is a probability distribution over source categories $\mathcal{Y}_s$. Then $p(y_s|y_t)$ can be calculated from $p(y_t|y_s)$ by Bayes's rule.

[1] https://github.com/makcedward/nlpaug

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBME.2022.3156293, IEEE Transactions on Biomedical Engineering

In addition, according to [29], it is necessary to calibrate the neural network, i.e. calibrating the probability output of the pre-trained model, to enhance performance.

3) Stochastic Normalization (StochNorm): In [30], stochastic normalization is proposed to avoid over-fitting during fine-tuning on small dataset. It replaces the Batch Normalization (BN) layers. It implements a two-branch architecture including one branch normalized by mini-batch statistics and another branch normalized by moving statistics (specified in detail below). A stochastic selection mechanism like Dropout is used between the two branches to avoid over-depending on some sample statistics. This is interpreted as an architecture regularization.

Let's assume a mini-batch of feature maps of each channel $z = \{z_i\}_{i=1}^m$ and a moving statistic update rate $\alpha \in (0, 1)$. The normalization process in the two branches is calculated as

$$\hat{z}_{i,0} = \frac{z_i - \tilde{\mu}}{\sqrt{\tilde{\sigma}^2 + \varepsilon}}, \quad \hat{z}_{i,1} = \frac{z_i - \mu}{\sqrt{\sigma^2 + \varepsilon}}, \tag{4}$$

during training, where the mean μ and variance σ^2 of the current mini-batch data of size m, i.e.

$$\mu \leftarrow \frac{1}{m}\sum_{i=1}^{m} z_i, \quad \sigma^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(z_i - \mu)^2 \tag{5}$$

are used as usual, while the other branch uses moving statistics $\tilde{\mu}$ and $\tilde{\sigma}^2$ of the training data

$$\tilde{\mu} \leftarrow \tilde{\mu} + \alpha(\mu - \tilde{\mu}), \quad \tilde{\sigma}^2 \leftarrow \tilde{\sigma}^2 + \alpha(\sigma^2 - \tilde{\sigma}^2). \tag{6}$$

The moving statistics are initialized by using the corresponding parameters from the pre-trained model[2]. During forward propagation, either $\hat{z}_{i,0}$ or $\hat{z}_{i,1}$ is randomly selected with probability p in each channel of the normalization layers and each training step, i.e.

$$\hat{z}_i = (1 - s)\hat{z}_{i,0} + s\hat{z}_{i,1}, \tag{7}$$

where s is the branch-selection variable generated from a Bernoulli distribution $s \sim Bernoulli(p)$. The learnable scale and shift parameters β, γ can be applied after the stochastic selection as usual

$$y_i \leftarrow \gamma \hat{x}_i + \beta. \tag{8}$$

Stochastic normalization in Fig. 1 uses a ResNet backbone where BN layers are replaced by StochNorm.

4) Combination of Co-tuning and Stochastic Normalization: We empirically evaluate the combination of co-tuning and StochNorm for lung sound classification. To do that, the category relationship is initially calculated based on the pre-trained ResNet models, followed by replacing BN layers with the StochNorm modules in the backbone of the original co-tuning architecture (i.e replacing the green ResBlocks from Co-tuning by the blue ResBlocks of Stochastic Normalization in Fig. 1). After that, co-tuning is processed on the new architecture.

IV. Experiments

In this section, we first provide details of the experimental setup. Furthermore, we empirically evaluate the following cases:

- Transfer learning of different pre-trained ImageNet ResNet models on the ICBHI dataset.
- Ablation study for respiratory segment length, spectrum corrections and flipping data augmentation.
- Transfer learning of different ResNet models pre-trained on ImageNet and ICBHI for our multi-channel lung sound dataset.

Our systems for the ALSC and RDC tasks on the ICBHI dataset are also compared against state-of-the-art works for the official ICBHI data split. Additionally, we compare our best system for crackle detection to our previous work on the multi-channel lung sound dataset.

A. Evaluation Metrics

We use the evaluation metrics supported by the ICBHI Challenge [5] for ALSC of 4 classes. The evaluation is based on respiratory cycles using sensitivity (*SE*), specificity (*SP*), average score (*AS*), known as the average of the sensitivity and the specificity, and the harmonic score (*HS*), known as the harmonic mean of the sensitivity and the specificity. For 2 classes, we determine *SE* and *SP* as in [22] and [17] and *AS* and *HS* as in [5].

Similarly, for RDC of 3 classes and 2 classes, a recording-wise evaluation is performed using *SE* and *SP* as in [22] and [17] and *AS* and *HS* as in [5].

Furthermore, for our multi-channel lung sound dataset, we calculate Precision (P_+), Sensitivity or Recall (Se), and the F-score as specified in [21]. Precision provides information about how many of the respiratory cycles recognized as crackles are actually true. Sensitivity provides information about how many respiratory cycles containing crackles are actually recognized as crackles. The F-score is the harmonic mean of precision and sensitivity.

We provide the 99% confidence interval (CI) for the average score and F-score in all bar charts and tables for the ICBHI dataset and our multi-channel lung sound dataset, respectively. The CI is computed from standard deviation over runs [39].

B. Experimental Setup

We evaluate our ALSC system for 4 and 2 classes and RDC task of 3 and 2 classes on the official ICBHI 2017 dataset split, which consists of 60% recordings for the training set and 40% for the test set. Each patient is either in the training or test set. The reported performance is the average score of five independent runs. As co-tuning requires a validation set to compute the category relationship, we randomly select 20% of the samples from the training set.

Due to the limited amount of data samples in our multi-channel lung sound dataset, we use 7-fold cross-validation with the recordings of each IPF subject appearing once in the test set. Each subject is assigned to either training, validation or test set. The best model is selected based on the best

[2] This exploits prior knowledge of pre-trained networks.

accuracy on the validation set. The reported performance of the system is an average accuracy of seven folds using the same data splittings.

Experiments are implemented based on Pytorch [40]. For vanilla fine-tuning, the learning rate and number of epochs is set to 0.001 and 150 for all tests, respectively. The fine-tuning of co-tuning and stochastic normalization techniques[3] updates the weights after each mini-batch. The learning rate of the feature representation layers and the last layer are set to 0.001 and 0.01, respectively. The fine-tuning process optimizes the cross entropy loss using SGD with a momentum of 0.9. The batch size is 32 for all experiments.

C. Experimental Results

1) Transfer learning techniques for different ResNets: We evaluate the vanilla fine-tuning (*VanillaFineTuning*), co-tuning (*CoTuning*), stochastic normalization (*StochNorm*) and the combination of co-tuning and stochastic normalization (*CoTuning-StochNorm*) for different ResNet architectures trained on the ImageNet dataset for the ALSC task of 4 classes (see Fig. 2) and the RDC task of 3 classes (see Fig. 4) on the official ICBHI dataset split. These systems use a segment length of 8s, spectrum correction using reference data s_{ref} of all devices and all data augmentation methods introduced in Section C.

Fig. 2 shows that ResNet50 is the best performing architecture to build the backbone for these transfer learning techniques of the 4-class ALSC task. ResNet101 is also performing well except for vanilla fine-tuning. Co-tuning achieved the best performance of ∼58% compared to the other techniques. Although *CoTuning* and *StochNorm* improve significantly the performance of *VanillaFineTuning*, the combination of co-tuning and StochNorm is not able to outperform the respective techniques for this task.

Fig. 3 visualizes the average pooling outputs of the ResNet50 architecture for different transfer learning techniques projected to 2D by t-distributed stochastic neighbourhood embedding (t-SNE) [41]. Distributions of the training set using vanilla fine-tuning, co-tuning, stochastic normalization and combination of co-tuning and stochastic normalization are shown in (a), (b), (c) and (d), respectively. Comparing to vanilla fine-tuning (a) and the stochastic normalization technique (c), the distribution of the 4 classes using co-tuning (b) and the combination of co-tuning and stochastic normalization (d) is better separated. It shows that the collaborative fine-tuning using the category relationship of source and target domain is useful for the adventitious lung sound classification task.

In Fig. 4, we see that the different transfer learning techniques using the ResNet101 model achieve the best performance for the 3-class RDC task. The ResNet50 model works better than the others for vanilla fine-tuning. *CoTuning-StochNorm* and *StochNorm* achieved a better performance compared to *CoTuning* and *VanillaFineTuning*. It proves the efficiency of the stochastic normalization in the fine-tuning process for the RDC task.

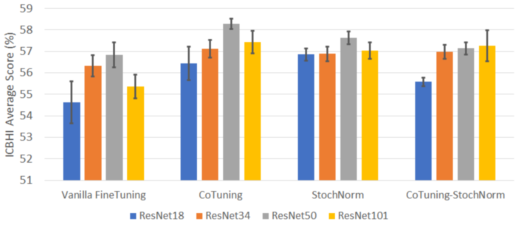

Fig. 2. Comparison of vanilla transfer learning, co-tuning, stochastic normalization and both co-tuning and stochastic normalization of different ResNet backbones for the adventitious lung sound classification task of 4 classes.

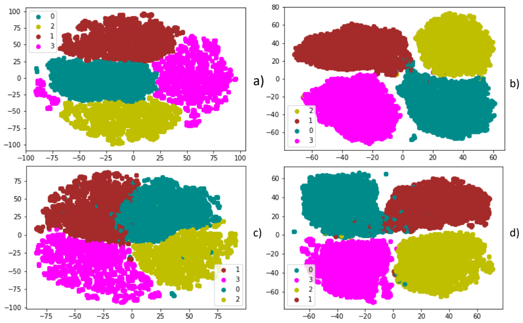

Fig. 3. Average pooling output representations reduced into 2D by t-distributed stochastic neighbourhood embedding (t-SNE) of the ResNet50 backbone architecture. Distributions of training set of (a) vanilla fine-tuning, (b) co-tuning, (c) stochastic normalization (d) combination of co-tuning and stochastic normalization. The color indicates the classes.

2) Respiratory segment length: The length of respiratory cycles in the ICBHI dataset varies in a wide range. Hence, we applied cycle splitting into segments and perform sample padding in order to obtain fixed-length segments. We observe different segment lengths for the ResNet 50 model fine-tuned by co-tuning and applied data augmentation in both time domain and time-frequency domain with spectrum correction. Results are shown in Table 3. The best AS is obtained with 8s fixed-length segments for 4 classes. We also use 8s segments for the other tasks of the ICBHI and our lung sound dataset.

3) Spectrum correction: We performed experiments on the ICBHI dataset with/without spectrum correction (calibration) using different reference spectra s_{ref}, which are determined by one or more devices. *No-Calib* denotes that no spectrum correction is applied. *Calib-Dev1* and *Calib-Dev2* denote calibration using data of device AKGC417L and Meditron, respectively. *Calib-Dev1Dev2* denotes calibration using data of both devices and *Calib-AllDev* denotes spectrum adaptation using reference data of all four devices. From Table II, we can see that co-tuning of the ResNet50 model using reference data of all devices achieves the best performance. It is 1.62% (absolute) better than without using spectrum calibration. Thus, we apply spectrum calibration for both adventitious lung sound classification and respiratory diseases classification.

4) Flipping data augmentation: We apply data augmentation in time domain and VTLP in order to balance the dataset. In this section we focus on the influence of feature flipping data augmentation (see Section III). Fig. 5 shows that when

[3] Code are available at https://github.com/thuml/Cotuning and https://github.com/thuml/StochNorm.

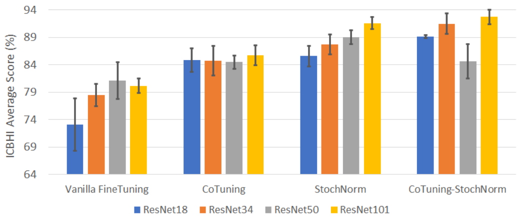

Fig. 4. Comparison of vanilla transfer learning, co-tuning, stochastic normalization and both co-tuning and stochastic normalization of different ResNet backbones for the respiratory disease classification task of 3 classes.

TABLE I
RESPIRATORY SEGMENT LENGTH: AVERAGE SCORE (AS) FOR VARIOUS INPUT LENGTH SIZES USING CO-TUNING OF RESNET50 AS BACKBONE NETWORK AND DATA AUGMENTATION WITHOUT SPECTRUM CALIBRATION.

Length.	**4 sec**	**5 sec**	**6 sec**	**7 sec**	**8 sec**	**9 sec**
AS±CI(%)	54.22 ± 1.18	56.25 ± 0.89	56.13 ± 0.61	56.55 ± 0.47	**56.67** **± 0.60**	56.58 ± 0.54

TABLE II
COMPARISON OF SPECTRUM CORRECTION METHODS USING DIFFERENT REFERENCE DATA. CO-TUNING OF THE RESNET50 WITH DATA AUGMENTATION IS USED.

	No-Calib	**Calib-Dev1**	**Calib-Dev2**	**Calib-Dev1Dev2**	**Calib-AllDev**
AS±CI(%)	56.67 ± 0.60	57.85 ± 0.21	57.85 ± 0.52	57.34 ± 0.14	**58.29** **± 0.24**

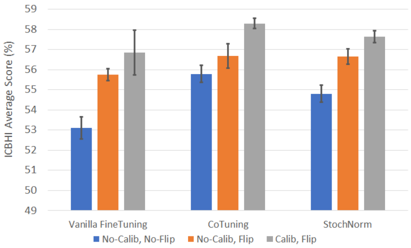

Fig. 5. Comparison of vanilla transfer learning, co-tuning and stochastic normalization using ResNet50 for the cases of spectrum correction and flipping data augmentation.

the system does not use spectrum correction, doubling the size of the augmented training set by the flipping technique always performs well for vanilla fine-tuning, co-tuning and stochastic normalization. It improves significantly the performance of vanilla fine-tuning and stochastic normalization of about 3% and 2%, respectively. For co-tuning, the flipping data augmentation achieves an improvement of 1% accuracy. Furthermore, we can see from Fig. 5 that using the combination of spectrum calibration and flipping data augmentation always enhances the robustness of the adventitious lung sound classification systems as the confidence intervals decrease.

5) Effect of pre-trained model on the multi-channel lung sound dataset: According to the above evaluation of transfer learning techniques for different residual neural networks for the 4-class ALSC task, co-tuning achieves the best performance. Thus, we evaluate the effect of pre-trained models using co-tuning (*CoTuning*) for the 2-class ALSC task on our multi-channel lung sound dataset. It is shown in Fig. 6. Co-tuning using the ImageNet pre-trained model always slightly outperforms that of the ICBHI pre-trained model. The smaller ResNet architectures tend to work better for co-tuning. We also can see from Fig. 6 that the ResNet34 backbone system achieves the best performance, followed by ResNet18, ResNet50 and ResNet101. In addition, Fig. 6 shows that transferred knowledge from full pre-trained models of ICBHI and the ImageNet dataset by co-tuning to our small lung sound dataset can achieve better accuracy than vanilla fine-tuning (*VanillaFineTuning*) using the ImageNet pre-trained model.

Overall, the best segment length for lung sound classification is 8s. Spectrum correction is useful to improve the performance of our ALSC and RDC system on the ICBHI dataset. This helps to correct the different frequency responses of the recording devices. The ALSC system using the flipping data augmentation enhances performance on both ICBHI and our multi-channel lung sound dataset. The new transfer learning methods always outperform vanilla transfer learning. Co-tuning works better for the ALSC task while StochNorm and its combination with co-tuning achieve higher performance for the RDC task. Furthermore, ResNet34 and ResNet50 are more suitable for the ALSC tasks, while a large ResNet101 model tends to be more robust for the RDC task in most transfer learning settings.

D. Performance Comparison

1) Comparison to state-of-the-art systems using the ICBHI dataset: Table IV and Table V show the comparison of our best systems of different transfer learning techniques and state-of-the-art systems (see Section V for more details on the systems) for the ALSC and RDC tasks, respectively. Our best systems are presented in bold and the highest scores are presented in bold and italic. It is notable that the performances on the official 60/40 ICBHI separation without common patients in both sets are significantly lower than that of randomly 80/20 splitting i.e. 5-fold cross validation and overlap of the same patients in both sets. Despite of the same fixed length for segments, the RDC systems always achieve considerably higher performance compared to the ALSC system for different sub-tasks. The RDC tasks have the full audio recordings which consists of many available respiratory cycles, while the ALSC tasks are processed and evaluated on respiratory cycles.

We evaluate our proposed system on the official ICBHI split for the 4 and 2 class ALSC tasks. Our best systems of different fine-tuning techniques outperform the other ALSC systems. Our system using co-tuning of the ResNet50 pre-trained model achieve the highest ICBHI average score of 58.29% and 64.74% for the 4-class and 2-class ALSC task, respectively. Our RDC systems are evaluated on the official ICBHI split of the 3-class RDC task, our systems achieves the best performance with the ResNet101 pre-trained architecture combined with stochastic normalization. It obtains 92.72% of the ICBHI average score. While on the 2-class RDC task, our systems using stochastic normalization achieves the average

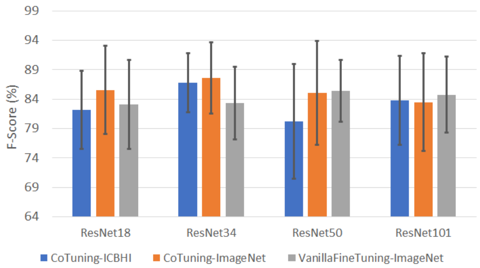

Fig. 6. Comparison of co-tuning and vanilla fine-tuning of different ResNet architectures pre-trained on ICBHI and ImageNet for crackle detection. Our multi-channel lung sound dataset and flipping data augmentation are used.

TABLE III
COMPARISON BETWEEN THE PROPOSED SYSTEMS AND THE SYSTEM IN [35] USING OUR MULTI-CHANNEL LUNG SOUND DATASET FOR CRACKLE DETECTION TASK.

Method	**Se(%)**	**P+(%)**	**F-Score±CI(%)**
MI-CNN, ICBHI FineTuning [35]	85.32	84.11	84.71
ResNet34, ICBHI CoTuning (Ours)	81.42	94.70	86.81±5.03
ResNet34, ImageNet CoTuning (Ours)	**82.17**	**95.88**	**87.59±6.04**
ResNet50, ImageNet StochNorm (Ours)	81.82	95.22	87.42±5.35
ResNet50, ImageNet CoTuning-StochNorm (Ours)	80.84	94.90	86.46±5.72
ResNet50, ImageNet VanillaFineTuning (Ours)	77.42	96.59	85.39±5.26

scores of 93.77% for the official splitting. Our best 2-class RDC system outperforms all compared systems.

2) Comparison for our multi-channel lung sound dataset: Table III compares our best systems using different transfer learning techniques with our previous system using fine-tuning for a multi-input CNN model [35] on the multi-channel lung sound dataset. We can see that our best transfer learning systems outperform the previous system. The co-tuning system using the ResNet34 model pre-trained on ImageNet achieves the best performance, closely followed by the StochNorm system using the pre-train ResNet50 model. The best F-score is 2.82% better than for the multi-input fine-tuned system [35].

V. RELATED WORKS

We review recent works on ALSC and RDC using the ICBHI 2017 dataset and binary ALSC works (i.e. crackle detection) using the multi-channel lung sound database. In general, it is difficult to compare the score of some proposed methods for ICBHI as substantial work does not use the official data splitting or use a different evaluation metric.

A. Lung Sound Classification on ICBHI dataset

There are two main directions: (i) conventional classifiers for low-level features of time or frequency domain, (ii) deep neural networks and robust machine learning techniques for spectral features.

1) Conventional approaches: Almost all proposed lung sound classification systems of the ICBHI challenge 2017 used conventional classifiers. The robust systems used hidden Markov models and Gaussian mixture models for MFCC features [50] or support vector machines (SVMs) for STFT and wavelet features [51].

Recently, a binary RDC system using the RUSBoost algorithm, which combines random under sampling and boosting techniques (i.e decision tree as a base classifier) was also introduced [49]. The input of the classifier are features selected from MFCCs, discrete wavelet transform and time domain features. The proposed system was evaluated on their own ICBHI dataset split and achieved 87.1% of average score.

In addition, Mukherjee et al. [48] developed a method to detect patients with respiratory infections. They extracted features based on linear predictive coefficient for a multilayer perceptron classifier. The method was evaluated on the ICBHI dataset using 5-fold cross-validation and achieved 99.22% of accuracy for the 2-class ALSC task.

2) Deep learning approaches: Deep learning systems use CNNs, recurrent neural networks (RNNs) and hybrid architectures. They are combined with machine learning techniques such as data augmentation, ensemble methods and transfer learning to enhance robustness. For the RNN-based systems, Kochetov et al. [32] proposed a system using a noise making RNN and MFCC features to classify cycles of lung sounds into four categories. The performance was evaluated based on 5-fold cross-validation. It is the first work which considers the effect of the recording devices on the performance. They achieved a score of 64.8% and 68.5% with training data from all devices and the most often occurring recording device (i.e. AKGC417L), respectively. In [22], Perna et. al. also introduced different architectures of RNNs such as long short time memory (LSTM), gated recurrent units (GRU), bidirectional-LSTM (BiLSTM) and bidirectional-GRU (BiGRU) for MFCC features to perform 4-class and 2-class ALSC and the 3- and 2-class RDC task. The results on random train-test ICBHI split of 80% and 20% are 74% and 81% of average score for the 4-class and 2-class ALSC task, respectively. The average performance of the RDC tasks of 3 and 2 classes is 84% and 91%, respectively.

Furthermore, CNNs or hybrid architectures have been used. In [24], we proposed a lung sound classification using a snapshot ensemble of CNNs for log-mel spectrograms. We applied temporal stretching and vocal tract length perturbation (VTLP) for data augmentation to deal with the class-imbalance of the ICBHI dataset. Our system achieved 78.4% and 83.7% of average score on the random train-test split of 80% and 20% with common patients in both sets for the ALSC task of 4 classes and 2 classes, respectively. Acharya et al. [27] introduced a deep CNN-RNN model for mel spectrograms to classify adventitious lung sounds into four classes. The performance for 5-fold cross validation evaluation was 66.31%. When this system was combined with a patient specific model tuning strategy, its performance increased up to 71.81% of average score. Similarly, Pham et al. [47], [17] introduced lung sound classification systems for adventitious sounds and respiratory diseases. In [47], they proposed various deep learning architectures mainly based on CNNs and RNNs using gammatone filtered spectrograms. They use a 80%-20% dataset split, where data from one subject may exist in both training and test set. An average ensemble of these systems achieved 80% and 86% average score for the 4-class and 2-class ALSC, respectively. The proposed CNN -

TABLE IV
COMPARISON BETWEEN THE PROPOSED SYSTEMS AND STATE-OF-THE-ART SYSTEMS FOR ALSC TASKS OF 4-CLASS AND 2-CLASS.

Task	**Method**	**Train/Test**	**SP(%)**	**SE(%)**	**AS±CI(%)**	**HS(%)**
ALSC, 4-class	STFT+Wavelet Spectrogram, CNN [42]	official 60/40	81	28	54	42
ALSC, 4-class	STFT+Wavelet Spectrogram, Bi-ResNet [43]	official 60/40	69.20	31.12	52.79	-
ALSC, 4-class	Gamatone Spectrogram, CNN-MoE, DA [17]	official 60/40	68	26	47	-
ALSC, 4-class	STFT Spectrogram, ResNet-NonLocal, DA [44]	official 60/40	63.20	41.32	52.26	-
ALSC, 4-class	STFT Spectrogram, ResNet-SE-SA [45]	official 60/40	81.25	17.84	49.55	-
ALSC, 4-class	STFT + Q-wavelet Spectrogram, ResNet-Attention [46]	official 60/40	71.44	33.15	53.90	-
ALSC, 4-class	Logmel Spectrogram, ResNet-FC, DA, BRC, Device Fine-tuning [33]	official 60/40	72.30	40.10	56.20	-
ALSC, 4-class	**Logmel Spectrogram, ResNet50, DA, SC, Vanilla Fine-tuning (Ours)**	**official 60/40**	**76.33**	**37.37**	**56.85±0.58**	**50.11**
ALSC, 4-class	**Logmel Spectrogram, ResNet50, DA, SC, StochNorm (Ours)**	**official 60/40**	**78.86**	**36.40**	**57.63±0.29**	**49.61**
ALSC, 4-class	**Logmel Spectrogram, ResNet50, DA, SC, CoTuning (Ours)**	**official 60/40**	***79.34***	***37.24***	***58.29±0.24***	***50.58***
ALSC, 4-class	**Logmel Spectrogram, ResNet101, DA, SC, CoTuning-StochNorm (Ours)**	**official 60/40**	**78.55**	**35.97**	**57.26±0.72**	**49.27**
ALSC, 4-class	MFCCs, NMRNN, Device Training [32]	5 folds	75	62	68.5	-
ALSC, 4-class	Mel Spectrogram, CNN-RNN. [27]	5 folds	84.14	48.63	66.38	-
ALSC, 4-class	STFT Spectrogram, ResNet-NonLocal, DA [44]	5 folds	64.73	63.69	64.21	-
ALSC, 4-class	Logmel Spectrogram, ResNet-FC, DA, BRC, Device Fine-tuning [33]	5 folds	83.30	53.70	68.50	-
ALSC, 4-class	MFCCs, LSTM [22]	overlap 80/20	85	62	74	-
ALSC, 4-class	Logmel Spectrogram, CNN Snapshot Ensembles, DA [24]	overlap 80/20	87.30	69.40	78.40	-
ALSC, 4-class	Gamatone Spectrogram, Ensemble, DA [47]	overlap 80/20	86	73	80	-
ALSC, 2-class	Gamatone Spectrogram, CNN-MoE, DA [17]	overlap 80/20	-	-	78.6	-
ALSC, 2-class	STFT + Q-wavelet Spectrogram, ResNet-Attention [46]	official 60/40	71.44	51.40	61.42	-
ALSC, 2-class	**Logmel Spectrogram, ResNet50, DA, SC, Vanilla Fine-tuning (Ours)**	**official 60/40**	**76.33**	**52.12**	**64.22±0.63**	**61.87**
ALSC, 2-class	**Logmel Spectrogram, ResNet50, DA, SC, StochNorm (Ours)**	**official 60/40**	**78.86**	**49.79**	**64.32±0.57**	**60.69**
ALSC, 2-class	**Logmel Spectrogram, ResNet50, DA, SC, CoTuning (Ours)**	**official 60/40**	***79.34***	***50.14***	***64.74±0.05***	***61.30***
ALSC, 2-class	**Logmel Spectrogram, ResNet101, DA, SC, CoTuning-StochNorm (Ours)**	**official 60/40**	**78.56**	**48.67**	**63.61±0.47**	**59.98**
ALSC, 2-class	CNN [33]	5 folds	80.90	73.10	77.0	-
ALSC, 2-class	Logmel, ResNet-FC, DA, BRC, Device Fine-tuning [33]	5 folds	80.90	73.10	77.0	-
ALSC, 2-class	MFCCs, LSTM [22]	overlap 80/20	-	-	81	-
ALSC, 2-class	Logmel Spectrogram, CNN Snapshot Ensembles, DA [24]	overlap 80/20	87.30	80.10	83.70	-
ALSC, 4-class	Gamatone Spectrogram, Ensemble, DA [47]	overlap 80/20	86	85	86	-
ALSC, 2-class	Gamatone Spectrogram, CNN-MoE, DA [17]	overlap 80/20	-	-	84.0	-

TABLE V
COMPARISON BETWEEN THE PROPOSED SYSTEMS AND STATE-OF-THE-ART SYSTEMS FOR RDC TASK OF 3-CLASS AND 2-CLASS.

Task	**Method**	**Train/Test**	**SP(%)**	**SE(%)**	**AS±CI(%)**	**HS(%)**
RDC, 3-class	Gamatone Spectrogram, CNN-MoE, DA [17]	official 60/40	-	-	84.0	-
RDC, 3-class	**Logmel Spectrogram, ResNet34, DA, SC, Vanilla Fine-tuning (Ours)**	**official 60/40**	**65.88**	**87.47**	**76.68±3.10**	**74.48**
RDC, 3-class	**Logmel Spectrogram, ResNet101, DA, SC, StochNorm (Ours)**	**official 60/40**	**90.59**	**92.53**	**91.56±1.10**	**91.35**
RDC, 3-class	**Logmel Spectrogram, ResNet101, DA, SC, CoTuning (Ours)**	**official 60/40**	**81.18**	**90.22**	**85.70±1.82**	**85.23**
RDC, 3-class	**Logmel Spectrogram, ResNet101, DA, SC, CoTuning-StochNorm (Ours)**	**official 60/40**	***91.77***	***93.68***	***92.72±1.30***	***92.57***
RDC, 3-class	MFCCs, LSTM [22]	overlap 80/20	82	98	90	-
RDC, 3-class	Gamatone Spectrogram, CNN-MoE, DA [47]	overlap 80/20	83	96	90	-
RDC, 3-class	Gamatone Spectrogram, CNN-MoE, DA [17]	5 folds	86	95	91	-
RDC, 2-class	Gamatone Spectrogram, CNN-MoE, DA [17]	official 60/40	-	-	84.1	-
RDC, 2-class	**Logmel Spectrogram, ResNet50, DA, SC, Vanilla Fine-tuning (Ours)**	**official 60/40**	**69.41**	**96.92**	**83.17±1.68**	**80.68**
RDC, 2-class	**Logmel Spectrogram, ResNet101, DA, SC, StochNorm (Ours)**	**official 60/40**	**90.59**	**93.90**	**92.25±1.11**	**90.02**
RDC, 2-class	**Logmel Spectrogram, ResNet18, DA, SC, CoTuning (Ours)**	**official 60/40**	**84.71**	**95.06**	**89.88±1.37**	**87.12**
RDC, 2-class	**Logmel Spectrogram, ResNet101, DA, SC, CoTuning-StochNorm (Ours)**	**official 60/40**	***91.77***	***95.76***	***93.77±1.41***	***93.60***
RDC, 2-class	MFCCs, LSTM [22]	overlap 80/20	82	99	91	-
RDC, 2-class	Gamatone Spectrogram, CNN-MoE, DA [47]	overlap 80/20	83	99	91	-
RDC, 2-class	Gamatone Spectrogram, CNN-MoE, DA [17]	5 folds	86	98	92	-
RDC, 2-class	LPC, MLP classifier [48]	5 folds	-	-	99.22	-
RDC, 2-class	MFCCs, DWT, time domain features, RUSBoost-DT [49]	50/50	93	86	87.10	-

mixture of expert (MoE) model was suitable for the RDC task of 3 classes and 2 classes with a performance of 90% and 91%, respectively. In [17], they proposed a CNN-MoE neural network for different feature types i.e. MFCCs, log-mel, gammatone filter and constant-Q transform spectrogram. The gammatone filter spectrogram was suggested for ALSC, while log-mel spectrograms worked better for the RDC task. The average score of the 4-class ALSC task was 47% for the ICBHI official dataset split. For 5-fold cross-validation with data of the same patient in both sets, their performance was 78.6% and 84% for the ALSC task of 4 classes and 2 classes, respectively. On the 3-class RDC task they achieved 85% of average score on the ICBHI official dataset split and 91% on 5-fold cross-validation.

Recently, CNN-based systems from diverse architectures i.e. VGGNets, ResNets or their variations have been more and more introduced. Minami et al. [42] proposed a 4-class ALSC system using a VGG16 neural network pre-trained on the ImageNet for the combination of STFT spectrogram and scalogram. The performance was 54% of average score on the official ICBHI dataset.

Ma et al. proposed two ALSC systems for four classes [43], [44]. The first one used an improved Bi-ResNet deep learning architecture based on STFT and wavelet features. Another system used non-local block ResNet with mixup data augmentation for STFT spectrograms. The proposed systems achieved 50.16% and 52.26% on the official data split, respectively. The latter work [44] was also evaluated using 5-fold cross-validation and achieved an average score of 64.21%. Yang et al. [45] proposed a 4-class ALSC system combining the ResNet18 architecture with Squeeze-and-Excitation and spatial attention blocks using STFT spectrogram features. They obtained 49.55% of average score on the official ICBHI dataset split. Li et. al. [46] proposed a deep architecture integrating an attention mechanism into the ResNet blocks for multi-channel spectrograms based on Q-factor wavelet transform and short-time Fourier transform. The performance for 4-class and 2-class ALSC is 53.9% and 61.42%, respectively.

Ordas et al. [52] proposed a CNN model for RDC using a variational convolutional autoencoder for data augmentation to balance the dataset. The achieved performance is 99.3% F-Score for the 3-class RDC task and 99.0% F-Score for the 6-class RDC task on the dataset split of 80% and 20% for training and test set, respectively. However, the systems were evaluated on the augmented test set and the performances can not be directly compared to the systems reported in Table V. Shuvo et al. [53] introduced their lightweight CNN model for detecting respiratory diseases using hybrid scalogram-based features of empirical mode decomposition and continuous wavelet transform. The proposed system achieved 98.92% for the three-class chronic classification task and 98.70% for the multi-class disease classification task. Similarly, these results are evaluated on the augmented test set.

Demir et al. proposed a 4-class ALSC system using pre-trained models for STFT spectrograms converted into color images. In the first approach [26], the pre-trained model was used as feature extractor and combined to an SVM classifier. In the second approach [26], the pre-trained model was fine-tuned on the ICBHI dataset. They achieved 65.5% and 63.09% of accuracy for 10-fold cross-validation, respectively. In [54], they introduced a parallel pooling CNN model for deep feature extraction. It is combined with a linear discriminant analysis classifier and random subspace ensembles. The performance of the proposed system was 71.5% for 10-fold cross-valuation. However, the evaluation metrics are different.

Additionally, Gairola et. al. [33] proposed a RespireNet model based on ResNet34 and fully connected layers with a set of techniques i.e. device specific fine-tuning, concatenation-based augmentation, blank region clipping and smart padding to improve the accuracy. The average score for the 4-class ALSC task was 56.2% and 68.5% for the official ICBHI dataset split and 5-fold cross-validation, respectively. They also evaluated the proposed system for the ALSC task of two classes and obtained 77.0% accuracy on 5-fold cross-validation.

B. Lung Sound Classification on our multi-channel dataset

In [21], Messner et al. introduced an event detection approach with bidirectional gated recurrent neural networks (BiGRNNs) using MFCCs to identify crackles in respiratory cycles. The proposed system was evaluated on the first version of the multi-channel lung sound dataset including 10 lung-healthy subjects and 5 patients with IPF. The performance was an F-score of 72% on 5-fold cross-validation.

In [14], a classification framework using lung sound signals of all recording channels was introduced to identify healthy and pathological breathing cycles. Lung sounds of one breath cycle of all recording channels were first transformed into STFT spectrograms. Then, the spectrogram were stacked into one compact feature vector. These features were fed into a CNN-RNN model for classification. Its score was 92% for 7-fold cross-validation.

We proposed a multi-input CNN model based on transfer learning for the detection of crackles and normal sounds on the multi-channel lung sound classification dataset. In [35], the multi-input CNN model shares the same network architecture of the pre-trained CNN model trained on the ICBHI dataset for respiratory cycles and their corresponding respiratory phases. Our system achieved an F-score of 84.71% using 7-fold cross-validation.

VI. Conclusion

We propose robust fine-tuning approaches to classify adventitious lung sounds and recognize respiratory diseases from lung auscultation recordings using the ICBHI and our multi-channel lung sound datasets. Transferred knowledge of pre-trained models from different ResNet architectures are exploited by vanilla fine-tuning, co-tuning, stochastic normalization and the combination of co-tuning and stochastic normalization techniques. Furthermore, spectrum correction and flipping data augmentation are introduced to improve the robustness of our system. Empirically, our proposed systems outperform almost all state-of-the-art systems for adventitious lung sound and respiratory disease classification. In particular, we obtain 58.29±0.24% and 64.74±0.05% average score for the 4- and 2-class adventitious lung sound task, respectively. Similarly, for the 3- and 2-class respiratory disease classification task, we obtain 92.72±1.30% and 93.77±1.41% average score, respectively. In addition, we also evaluate our adventitious lung sound classification approach using co-tuning on our multi-channel lung sound dataset to detect crackles using different pre-trained models of the ImageNet and ICBHI dataset. The best co-tuning system for 2-class lung sound classification achieves a better F-score (2.82%) compared to our previous work using a multi-input convolutional neural network. We also review state-of-the-art classification systems for adventitious lung sounds and respiratory diseases using the ICBHI dataset and our multi-channel lung sound dataset.

Acknowledgment

The authors would like to thank our colleague, Alexander Fuchs for feedback and fruitful discussions.

References

[1] WHO, "https://covid19.who.int/," Accessed May 11, 2021.

[2] M.T. Barbosa et. al., "The "big five" lung diseases in covid-19 pandemic–a google trends analysis," *Pulmonology*, vol. 27, no. 1, pp. 71, 2021.

[3] N. Chen et. al., "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study," *The lancet*, vol. 395, no. 10223, pp. 507–513, 2020.

[4] M. Sarkar et. al., "Auscultation of the respiratory system," *Annals of thoracic medicine*, vol. 10, no. 3, pp. 158, 2015.

[5] B. M. Rocha et. al., "A respiratory sound database for the development of automated classification," in *Precision Medicine Powered by pHealth and Connected Health*, pp. 33–37. Springer, 2018.

[6] R. X. A. Pramono et. al., "Automatic adventitious respiratory sound analysis: A systematic review," *PloS one*, vol. 12, no. 5, pp. e0177926, 2017.

[7] M. Pahar et. al., "Covid-19 cough classification using machine learning and global smartphone recordings," *Computers in Biology and Medicine*, p. 104572, 2021.

[8] K. K. Lella et. al., "Automatic diagnosis of covid-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: Cough, voice, and breath," *Alexandria Engineering Journal*, 2021.

[9] Y. Chang et. al., "Covnet: A transfer learning framework for automatic covid-19 detection from crowd-sourced cough sounds," *Frontiers in Digital Health*, p. 195.

[10] L. Orlandic et.al., "The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021.

[11] H. H. Ying et. al., "The respiratory sound features of covid-19 patients fill gaps between clinical data and screening methods," *medRxiv*, 2020.

[12] E. A. Lapteva et. al., "Automated lung sound analysis using the lungpass platform: a sensitive and specific tool for identifying lower respiratory tract involvement in covid-19," *European Respiratory Journal*, 2021.

[13] Dataset:, "Rale: A computer-assisted instructional package," in *Respiratory Care*. 35, 1006, 1990.

[14] E. Messner et. al., "Multi-channel lung sound classification with convolutional recurrent neural networks," *Computers in Biology and Medicine*, p. 103831, 2020.

[15] M. Fraiwan et. al., "A dataset of lung sounds recorded from the chest wall using an electronic stethoscope," *Data in Brief*, vol. 35, pp. 106913, 2021.

[16] H. Chen et. al., "Triple-classification of respiratory sounds using optimized s-transform and deep residual networks," *IEEE Access*, vol. 7, pp. 32845–32852, 2019.

[17] L. D. Pham et. al., "Cnn-moe based framework for classification of respiratory anomalies and lung disease detection," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[18] L.P. Malmberg et. al., "Classification of lung sounds in patients with asthma, emphysema, fibrosing alveolitis and healthy lungs by using self-organizing maps," *Clinical Physiology*, vol. 16, no. 2, pp. 115–129, 1996.

[19] M. Bahoura, "Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes," *Computers in biology and medicine*, vol. 39, no. 9, pp. 824–843, 2009.

[20] P. Bokov et. al., "Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population," *Computers in biology and medicine*, vol. 70, pp. 40–50, 2016.

[21] E. Messner et. al., "Crackle and breathing phase detection in lung sounds with deep bidirectional gated recurrent neural networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 356–359.

[22] D. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2019, pp. 50–55.

[23] M. Aykanat et. al., "Classification of lung sounds using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 65, 2017.

[24] T. Nguyen and F. Pernkopf, "Lung sound classification using snapshot ensemble of convolutional neural networks," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 760–763.

[25] K. He et. al., "Rethinking imagenet pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4918–4927.

[26] F. Demir et. al., "Convolutional neural networks based efficient approach for classification of lung diseases," *Health information science and systems*, vol. 8, no. 1, pp. 1–8, 2020.

[27] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *IEEE transactions on biomedical circuits and systems*, vol. 14, no. 3, pp. 535–544, 2020.

[28] L. Shi et. al., "Lung sound recognition algorithm based on vggish-bigru," *IEEE Access*, vol. 7, pp. 139438–139449, 2019.

[29] K. You et. al., "Co-tuning for transfer learning," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[30] Z. Kou et.al., "Stochastic normalization," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[31] T. Nguyen et. al., "Acoustic scene classification for mismatched recording devices using heated-up softmax and spectrum correction," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 126–130.

[32] K. Kochetov et. al., "Noise masking recurrent neural network for respiratory sound classification," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 208–217.

[33] S. Gairola et. al., "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2021, pp. 527–530.

[34] A. M. Feldman, "Majority voting," in *Welfare Economics and Social Choice Theory*, pp. 196–215. Springer, 1989.

[35] T. Nguyen and F. Pernkopf, "Crackle detection in lung sounds using transfer learning and multi-input convolutional neural networks," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2021, pp. 80–83.

[36] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[37] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Applied Sciences*, vol. 6, no. 2, pp. 57, 2016.

[38] Na. Jaitly and G.E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, vol. 117.

[39] Handbook Of Parametric, "Handbook of parametric and nonparametric statistical procedures," .

[40] A. Paszke et. al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[41] L. V. D. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[42] K. Minami et. al., "Automatic classification of large-scale respiratory sound dataset based on convolutional neural network," in *2019 19th International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2019, pp. 804–807.

[43] Y. Ma et. al., "Lungbrn: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2019, pp. 1–4.

[44] Y. Ma et. al., "Lungrn+ nl: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation," in *Proc. Interspeech*, 2020, vol. 2020, pp. 2902–2906.

[45] Z. Yang et. al., "Adventitious respiratory classification using attentive residual neural networks," in *Proceedings of the Interspeech*, 2020, pp. 2912–2916.

[46] J. Li et. al., "Lungattn: advanced lung sound classification using attention mechanism with dual tqwt and triple stft spectrogram," *Physiological Measurement*, vol. 42, no. 10, pp. 105006, 2021.

[47] L. Pham et. al., "Robust deep learning framework for predicting respiratory anomalies and diseases," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 164–167.

[48] H. Mukherjee et. al., "Automatic lung health screening using respiratory sounds," *Journal of Medical Systems*, vol. 45, no. 2, pp. 1–9, 2021.

[49] X. H. Kok et. al., "A novel method for automatic identification of respiratory disease from acoustic recordings," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2589–2592.

[50] N. Jakovljević and T. Lončar-Turukalo, "Hidden markov model based respiratory sound classification," in *International Conference on Biomedical and Health Informatics*. Springer, 2017, pp. 39–43.

[51] G. Serbes et. al., "An automated lung sound preprocessing and classification system based onspectral analysis methods," in *International Conference on Biomedical and Health Informatics*. Springer, 2017, pp. 45–49.

[52] M. T. García-Ordás et. al., "Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data," *Sensors*, vol. 20, no. 4, pp. 1214, 2020.

[53] S. B. Shuvo et. al., "A lightweight cnn model for detecting respiratory diseases from lung auscultation sounds using emd-cwt-based hybrid scalogram," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[54] F. Demir et. al., "Classification of lung sounds with cnn model using parallel pooling structure," *IEEE Access*, vol. 8, pp. 105376–105383, 2020.