2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

RespireNet: A Deep Neural Network for Accurately Detecting Abnormal Lung Sounds in Limited Data Setting

Siddhartha Gairola[1], Francis Tom[2], Nipun Kwatra[1] and Mohit Jain[1]

***Abstract*— Auscultation of respiratory sounds is the primary tool for screening and diagnosing lung diseases. Automated analysis, coupled with digital stethoscopes, can play a crucial role in enabling tele-screening of fatal lung diseases. Deep neural networks (DNNs) have shown potential to solve such problems, and are an obvious choice. However, DNNs are data hungry, and the largest respiratory dataset ICBHI has only 6898 breathing cycles, which is quite small for training a satisfactory DNN model. In this work, *RespireNet*, we propose a simple CNN-based model, along with a suite of novel techniques—device specific fine-tuning, concatenation-based augmentation, blank region clipping, and smart padding—enabling us to efficiently use the small-sized dataset. We perform extensive evaluation on the ICBHI dataset, and improve upon the state-of-the-art results for 4-class classification by 2.2%.**

Code: `https://github.com/microsoft/RespireNet`

I. Introduction

Respiratory diseases like asthma, chronic obstructive pulmonary disease (COPD), lower respiratory tract infection, lung cancer, and tuberculosis are the leading causes of death worldwide [1]. Early diagnosis has been found to be crucial in effectively treating respiratory diseases and reducing their adverse effects on the length and quality of life. Listening to chest sounds using a stethoscope is a standard method for screening and diagnosing lung diseases. It provides a low-cost and non-invasive screening methodology, avoiding the exposure risks of radiography and patient-compliance requirements associated with tests such as Spirometry.

There are a few drawbacks of stethoscope-based diagnosis: requirement of a trained medical professional to interpret auscultation sounds, and subjectivity in interpretations causing inter-listener variability. These limitations are exacerbated in impoverished settings and during pandemic situations (such as COVID-19), due to shortage of expert medical professionals. Automated analysis of respiratory sounds can help in alleviating these drawbacks, and also help in enabling tele-medicine applications to monitor patients outside a clinic by less-skilled workforce such as community health workers.

Algorithmic detection of lung diseases from respiratory sounds has been an active area of research [2, 3], especially with the advent of digital stethoscopes. Most of these works focus on detecting abnormal respiratory sounds of *wheeze* and *crackle*. Wheeze is a typical symptom of asthma and COPD, characterized by a high-pitched continuous sound in the frequency range of 100-2500Hz and duration above 80 msec [3, 4]. Crackles, which are associated with COPD, chronic bronchitis, pneumonia and lung fibrosis [5, 6], have a discontinuous, non-tonal sound, with frequency of ∼650 Hz and duration of 5 msec (for fine crackles), or frequency of 100-500 Hz and duration of 15 msec (for coarse crackles).

[1] Microsoft Research India. {t-sigai, nkwatra, mohja}@microsoft.com
[2] Microsoft India. francis.tom@microsoft.com

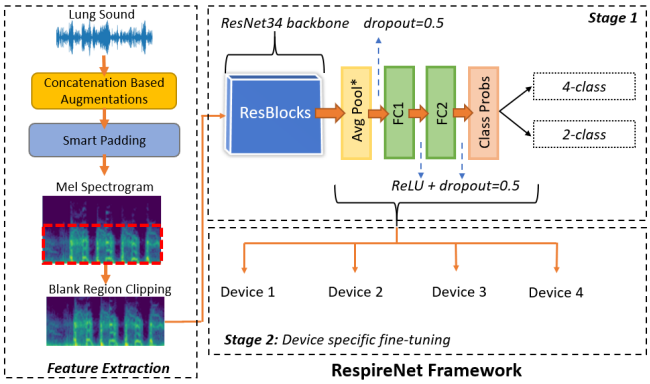

Fig. 1. *Overview of proposed RespireNet framework. We pre-process the audio signal (bandpass filtering, downsampling, normalization, etc.), apply concatenation-based augmentation and smart padding, and generate the Mel-spectrogram. Blank region clipping is applied to remove blank regions in the high frequency ranges. The processed spectrogram is then used to train our DNN model via a two-stage training. Stage-1: the model is trained using entire train set. Stage-2: device specific fine-tuning which trains using subset of data corresponding to each device.*

Although early works on abnormal lung sounds detection focused on hand-crafted features and traditional machine learning [7, 8], more recently, deep learning based methods have been explored [9, 10, 11]. For training DNNs, a time-frequency representation of the audio signal, such as Mel-spectrograms [10, 12, 13], stacked MFCC features [9, 12, 14, 15, 16], or optimized S-transform spectrogram [17] has been used. This 2D "image" is then fed into CNNs [14, 15], RNNs [9, 18], or hybrid CNN-RNNs [10] to learn robust high dimensional representations.

It is well known that DNNs typically require large datasets to achieve good performance. In this work, we use the ICBHI respiratory sound challenge dataset [19]. Despite being the largest publicly available dataset, it has only 6898 breathing cycle samples, which is quite small for training deep networks. Thus, a big focus of our work has been on developing a suite of techniques to help train DNNs in a data efficient manner. For that, we analyzed the ICBHI dataset extensively, and found several characteristics of the data that might inhibit training DNNs effectively. For example, the dataset contains audio recordings from four different devices, with skewed distribution of samples across the devices. Similarly, the dataset has a skewed distribution across normal and abnormal classes, and varying lengths of audio samples.

We propose multiple novel techniques to address these problems—device specific fine-tuning, concatenation-based augmentation, blank region clipping, and smart padding.

Combined with these techniques, we found that a simple CNN architecture, such as ResNet, was able to achieve high accuracy. This is in contrast to prior work employing complex architectures, like hybrid CNN-RNN [10], non-local block additions to CNNs [11], etc. Finally we perform extensive evaluation and ablation analysis of our model, *RespireNet*, improving upon the state-of-the-art results for 4-class classification by 2.2%.

The main contributions of our work are:

1) demonstrate that a simple network architecture is sufficient for respiratory sound classification, and more focus is needed to make efficient use of available data.
2) a detailed analysis of the ICBHI dataset pointing out its characteristics impacting DNN training significantly.
3) a suite of techniques—device specific fine-tuning, concatenation-based augmentation, blank region clipping, smart padding—enabling efficient dataset usage. These techniques are independent of network architecture and can be easily incorporated in other networks.

II. Method

Dataset: We perform all evaluations on the ICBHI scientific challenge respiratory sound dataset [19]. The dataset comprises of 920 recordings from 126 patients with a combined total duration of 5.5 hours. Each breathing cycle in a recording is annotated by an expert as one of the four classes: *normal*, *crackle*, *wheeze*, or *both* (crackle and wheeze). The dataset comprises of recordings from four different devices[1] from hospitals in Portugal and Greece. For every patient, data was recorded at seven different body locations.

Pre-processing: The sampling rate of recordings in the dataset varies from 4 kHz to 44.1 kHz. To standardize, we downsample the recordings to 4 kHz, and apply a 5-th order Butterworth band-pass filter to preserve frequencies in the range of 50-2000 Hz and to remove noise (heartbeat sound, background speech, etc.). Since the audio is recorded using different devices in different environments, we apply standard normalization on the input signal to map the values within the range (-1, 1). The audio signal is then converted into a Mel-spectrogram (similar to [10, 12, 13]), which is fed into our DNN.

Network architecture: We use a CNN-based network (Figure 1), *ResNet-34*[2], followed by two *128-d* fully connected linear layers with *ReLU* activations. The last layer applies *softmax activation* to model classwise probabilities. Dropout is added to the fully-connected layers to prevent overfitting. The network is trained via a standard categorical cross-entropy loss to minimize the loss for multi-class classification.

A. Efficient Dataset Utilization

To efficiently use the available 6898 samples, we extensively analyzed the dataset to identify characteristics that inhibit training DNNs effectively, and propose solutions to overcome the same. The first commonly used technique we apply is *transfer learning*, where we initialize our network with weights of a pre-trained *ResNet-34* network on ImageNet [20]. This is followed by training the network end-to-end. Interestingly, even though ImageNet dataset is very different from our Mel-spectrograms, we still found this initialization to help significantly. Most likely, low level features such as edge-detection are still similar and thus "transfer" well.

Concatenation-based Augmentation: Like most medical datasets, ICBHI dataset has a huge class imbalance, with the *normal* class accounting for 53% of the samples. To prevent the model from overfitting, we experimented with several data augmentation techniques. We first apply standard audio augmentation techniques, such as noise addition, speed variation, random shifting, pitch shift, etc., and also use a weighted random sampler to sample mini-batches uniformly from each class. These standard techniques help a little, but to further improve generalization of the under-represented classes (*wheeze*, *crackle*, *both*), we developed a concatenation-based augmentation technique where we generate a new sample of a class by randomly sampling two samples of the same class and concatenating them (see Figure 2). This scheme led to a non-trivial improvement in the classification accuracy of abnormal classes.

Fig. 2. *Proposed concatenation-based augmentation.*

Smart Padding: The breathing cycle length varies across patients as well as within a patient due to various factors (*e.g.*, breathing rate can increase during fever). In the ICBHI dataset, the length of breathing cycles ranges from 0.2s to 16.2s (mean=2.7s). This poses a problem while training our network as it expects a fixed size input[3]. The standard way to handle this is to pad the audio signal to a fixed size via *zero-padding* or *reflection*-based padding. We propose a novel *smart padding* scheme, which uses a variant of our *augmentation* scheme. For each data sample, *smart padding* examines the breathing cycle sample for that patient taken just before and after the current one. If either of the neighbouring cycle is of the same class or of the *normal* class, we concatenate the current sample with it. If not, we pad by copying the same cycle again. We continue this process until we reach our desired size. This padding scheme also augments data and helps prevent overfitting.

Blank Region Clipping: On analyzing samples using Grad-Cam++ [21] which our base model misclassified, we found significant blank regions[4] at higher frequency regions of their spectrograms (Figure 3). On further analysis, we found that many samples had blank region in the 1.5-2kHz frequency range (*e.g.*, 20% of samples from the Meditron device had 1.5-2kHz frequency range missing, while all Litt3200 device samples had the same 1.5-2kHz frequency missing).

[1]The four devices used for recordings are AKGC417L Microphone, 3M Littmann Classic II SE Stethoscope, 3M Litmmann 3200 Electronic Stethoscope, and WelchAllyn Meditron Master Elite Electronic Stethoscope.

[2]ResNet-18 performed poorly compared to ResNet-34, while ResNet-50 performed similar to ResNet-34.

[3]CNNs can be made size agnostic by using adaptive average pooling, however that typically hurts accuracy.

[4]Blank region in a spectrogram means that the audio signal has zero energy in the corresponding audio frequency range (Figure 3a).

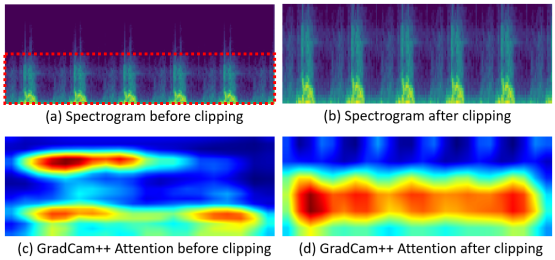

Fig. 3. *Blank region clipping: The network attention [21] starts focusing on the bottom half of the spectrogram, instead of blank spaces, after clipping.*

This causes a sharp discontinuity (edges) between 'data' region and these blank regions, which adversely impacts the CNN model, as the initial few layers are known to detect edges [22]. Since it was hurting our network performance, we selectively clip off the high frequency blank regions from the spectrograms. The blank region clipping has no impact on samples having information in the higher frequency ranges, and those samples are used as is. This ensures that (1) no useful information gets pruned, and (2) the CNN is not confused by the sudden discontinuity caused by the blank regions, thus improving performance.

Device Specific Fine-tuning: The ICBHI dataset has samples from 4 different devices. We found that the distribution of samples across devices is heavily skewed, *e.g.* 63% samples are recorded using the AKGC417L Mic (Table III). Since each device has different audio characteristics, the DNN may fail to generalize across devices, especially for the underrepresented devices in the already small dataset. To verify this, we divided the test set into 4 subsets based on the device type, and compute the accuracy of abnormal class samples in each subset. As expected, we found the classification accuracy to be strongly correlated with the training set size of the corresponding device. To address this, we first train a common model with the full training data (stage-1, Figure 1). We then make 4 copies of this model and *fine-tune* (stage-2) them for each device separately by using only the subset of training data for that device. We found this approach to significantly improve the performance, especially for the underrepresented devices (Table III).

III. Experiments

We evaluate the performance of our framework on the respiratory anomaly classification task proposed in the ICBHI challenge [19]. It consists of two subtasks: (i) classify a breathing cycle into one of the four classes–*normal(n), crackle(c), wheeze(w), both(b)*, and (ii) classify a breathing cycle into *normal or abnormal* class, where $abnormal = \{crackle, wheeze, both\}$. The final score is computed as the mean of Sensitivity: $S_e = \frac{P_c+P_w+P_b}{N_c+N_w+N_b}$ and Specificity: $S_p = \frac{P_n}{N_n}$, where P_i and N_i are the number of correctly classified and total number of samples in class i, respectively ($i \in \{normal, crackle, wheeze, both\}$). For the 2-class case, we adopt the abnormal and normal class scores as S_e and S_p respectively, and the score is computed as their mean.

We compare our performance using the above evaluation metric on two dataset divisions: the official 60-40% split [19] and 80-20% split [10, 11, 23] for train-test[5]. For the 4-class classification task, *RespireNet* outperforms state-of-the-art (SOTA) by 3.9% [11] on the official 60-40 split, and by 2.2% [10] on the 80-20 split (Table I). Further, *RespireNet* achieves the new SOTA of 77.0% on the 2-class task.

[5]For both the splits, the train and test set are patient-wise disjoint.

Split & Task	Method	S_p	S_e	Score
60-40 Split & 4-class	Jakovljevic et al. [7]	-	-	39.5%
	Chambres et al. [8]	78.1%	20.8%	49.4%
	Serbes et al. [24]	-	-	49.9%
	Ma et al. [23]	69.2%	31.1%	50.2%
	Ma, Xu, and Li [11]	63.2%	41.3%	52.3%
	CNN (ours)	71.4%	39.0%	55.2%
	CNN+CBA+BRC (ours)	71.8%	39.6%	55.7%
	CNN+CBA+BRC+FT (ours)	72.3%	40.1%	**56.2%**
80-20 Split & 4-class	Kochetov et al. [9]	73.0%	58.4%	65.7 %
	Acharya et al. [10]	84.1%	48.6%	66.3%
	Ma, Xu, and Li [11]	64.7%	63.7%	64.2%
	CNN (ours)	78.8%	53.6%	66.2%
	CNN+CBA+BRC (ours)	79.7%	54.4%	67.1%
	CNN+CBA+BRC+FT (ours)	83.3%	53.7%	**68.5%**
80-20 Split & 2-class	CNN (ours)	83.3%	60.5%	71.9%
	CNN+CBA+BRC (ours)	76.4%	71.0%	73.7%
	CNN+CBA+BRC+FT (ours)	80.9%	73.1%	**77.0%**

TABLE I

Performance comparison of our proposed model with SOTA on different splits. Our proposed techniques – concatenation-based augmentation (CBA), blank region clipping (BRC) and device specific fine-tuning (FT) – led to significant improvements.

Length	1 sec	2 sec	3 sec	4 sec	5 sec	6 sec	7 sec	8 sec	9 sec
Scores	56.6	59.0	60.3	61.1	62.3	64.4	**66.2**	65.1	65.5

TABLE II

Breathing cycle length size versus classification score.

Implementation Details: We train our model on a Tesla v100 GPU on a Microsoft Azure VM. We used the SGD optimizer with momentum of 0.9, and a batch size of 64. We used a fixed learning rate of 1e-3 for stage-1 and 1e-4 for stage-2 of training. Stage-1 was trained for 200 epochs. The highest validation checkpoint from stage-1 was used to train stage-2 for another 50 epochs for each device.

We further analyze the effect of our novel proposed techniques by conducting an ablation analysis on the 4-class classification task on the 80-20 split.

Concatenation-based Augmentation: Due to the small number of abnormal samples, our CNN model tends to overfit[6] on the abnormal classes, and achieved a score of 62.2%. Standard augmentations (noise addition, etc.) improved the score to 66.2%, which further improved to 66.8% with our concatenation-based augmentation. Also, this gain was mainly due to improved accuracy of the abnormal classes, where the sensitivity increased by 1.5%. This shows that our augmentation scheme to generate novel samples for the abnormal classes help the model generalize better.

Smart Padding: We experimented with different breathing cycle lengths and found 7s length to perform best (Table II). A small cycle length led to clipping of samples, thus loosing valuable information in an already scarce dataset, while a big cycle length caused repetition leading to degraded performance. For the base model, *smart padding* improves

[6]The generalization gap (test error – train error) was much higher for the abnormal classes compared to the normal class, despite the low train error for both classes. On continuing our training for longer, we observed that the test error for the abnormal classes started increasing.

Device	AKGC417L	Meditron	Litt3200	LittC2SE
% Samples	63%	21%	9%	7%
Score Impr	1.7%	1.6%	9.3%	8.6%

TABLE III

Device specific fine-tuning: Devices with small number of samples show a big score improvement.

accuracy over *zero-padding* and *reflection-based* padding by 5% and 2% respectively. This demonstrates the effectiveness of our padding scheme.

Blank Region Clipping: It resulted in an improvement of 0.5% over the base model score of 66.2%. When combined with our proposed augmentation, it helped achieve a score of 67.1%, outperforming the current SOTA [10] by 0.8%.

Device specific fine-tuning: Our fine-tuning resulted in an improvement of 1.4% in the final ICBHI score. It disproportionally helped the under-represented classes; devices with fewer samples had $\sim$9% increase in their scores (Table III).

IV. Related Work

Recently, there has been a lot of interest in using deep learning models for respiratory sounds classification [10, 11, 9]. It has outperformed statistical methods (HMM-GMM) [7] and traditional machine learning methods (boosted decision trees, SVM) [8, 24]. In these DNNs, a time-frequency representation of the audio signal is provided as input to the model. Kochetov et al. [9] propose a deep RNN with a noise masking intermediate step for the 4-class classification task, obtaining a score of 65.7% on the 80-20 split. However the paper omits detail about noise label generation [10], thus making it hard to reproduce. Deep residual networks and optimized S-transform based features are used by Chen et al. [17] for three-class anomaly classification in lung sounds. The model is trained and tested on a smaller subset of the ICBHI dataset on a 70-30 split and achieve a score of 98%.

Acharya and Basu [10] propose a Mel-spectrogram based hybrid CNN-RNN model with patient-specific model tuning, achieving a score of 66.3% on 4-class and 80-20 split. Ma, Xu, and Li [11] introduce LungRN+NL which incorporates a non-local block in the ResNet architecture and apply mixup augmentations to address the data imbalance problem, achieving sensitivity of 63.7%. However, none of these approaches focus on the audio characteristics of the ICBHI dataset, which we exploit to improve performance.

V. Conclusion and Future Work

The paper proposes *RespireNet*, a simple CNN-based model, along with a set of novel techniques—device specific fine-tuning, concatenation-based augmentation, blank region clipping, and smart padding—enabling effective utilization of a small-sized dataset for accurate abnormality detection in lung sounds. Our proposed method achieved a new SOTA for the ICBHI dataset, on both the 2-class and 4-class classification tasks. Further, our proposed techniques are independent of the choice of network architecture and should be easy to incorporate within other frameworks.

The current performance limit of the 4-class classification task can be mainly attributed to the small size of the ICBHI dataset, and the variation among the recording devices. Furthermore, there is lack of standardization in the 80-20 split and we found variance in the results based on the particular split. In future, we would recommend that the community should focus on capturing a larger dataset, while taking care of the issues raised in this paper.

References

[1] WHO. "The global impact of respiratory diseases (2nd edition)". In: *Forum of International Respiratory Societies (FIRS)* (2017).

[2] Hüseyin Polat and Inan Guler. "A Simple Computer-Based Measurement and Analysis System of Pulmonary Auscultation Sounds". In: *Journal of medical systems* 28 (Jan. 2005), pp. 665–72.

[3] Sandra Reichert et al. "Analysis of Respiratory Sounds: State of the Art". In: *Clinical Medicine : Circulatory, Respiratory and Pulmonary Medicine* 2 (May 2008).

[4] Abraham Bohadana, Gabriel Izbicki, and Steve Kraman. "Fundamentals of Lung Auscultation". In: *The New England journal of medicine* 370 (Feb. 2014), pp. 744–751.

[5] B Flietstra et al. "Automated Analysis of Crackles in Patients with Interstitial Pulmonary Fibrosis". In: *Pulmonary medicine* 2011 (Jan. 2011), p. 590506.

[6] Renard Xaviero Adhi Pramono, Stuart A. Bowyer, and E. Rodríguez-Villegas. "Automatic adventitious respiratory sound analysis: A systematic review". In: *PLoS ONE* 12 (2017).

[7] Niksa Jakovljevic and Tatjana Loncar-Turukalo. "Hidden Markov Model Based Respiratory Sound Classification". In: Jan. 2018, pp. 39–43.

[8] Gaetan Chambres, Pierre Hanna, and Myriam Desainte-Catherine. "Automatic Detection of Patient with Respiratory Diseases Using Lung Sound Analysis". In: Sept. 2018, pp. 1–6.

[9] Kirill Kochetov et al. "Noise Masking Recurrent Neural Network for Respiratory Sound Classification". In: *Artificial Neural Networks and Machine Learning*. Oct. 2018, pp. 208–217.

[10] Jyotibdha Acharya and Arindam Basu. "Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning". In: *IEEE Transactions on Biomedical Circuits and Systems* PP (Mar. 2020), pp. 1–1.

[11] Yi Ma, Xinzi Xu, and Yongfu Li. "LungRN+NL: An Improved Adventitious Lung Sound Classification Using non-local block ResNet Neural Network with Mixup Data Augmentation". In: Aug. 2020.

[12] Lukui Shi et al. "Lung Sound Recognition Algorithm Based on VGGish-BiGRU". In: *IEEE Access* PP (Sept. 2019), pp. 1–1.

[13] Renyu Liu et al. "Detection of Adventitious Respiratory Sounds based on Convolutional Neural Network". In: 2019, pp. 298–303.

[14] Murat Aykanat et al. "Classification of lung sounds using convolutional neural networks". In: *EURASIP Journal on Image and Video Processing* (2017), pp. 1–9.

[15] Diego Perna. "Convolutional Neural Networks Learning from Respiratory data". In: Dec. 2018, pp. 2109–2113.

[16] Elmar Messner et al. "Crackle and Breathing Phase Detection in Lung Sounds with Deep Bidirectional Gated Recurrent Neural Networks". In: vol. 2018. July 2018, pp. 356–359.

[17] Hai Chen et al. "Triple-Classification of Respiratory Sounds Using Optimized S-Transform and Deep Residual Networks". In: *IEEE Access* PP (Mar. 2019), pp. 1–1.

[18] 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)* (2019), pp. 50–55.

[19] B. M. Rocha et al. "A Respiratory Sound Database for the Development of Automated Classification". In: *Precision Medicine Powered by pHealth and Connected Health*. Singapore: Springer, 2018, pp. 33–37.

[20] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252.

[21] A. Chattopadhyay et al. "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks". In: *IEEE WACV* (2018), pp. 839–847.

[22] Matthew D. Zeiler and R. Fergus. "Visualizing and Understanding Convolutional Networks". In: *ECCV*. 2014.

[23] Yi Ma et al. "LungBRN: A Smart Digital Stethoscope for Detecting Respiratory Disease Using bi-ResNet Deep Learning Algorithm". In: *2019 IEEE BioCAS*. Oct. 2019, pp. 1–4.

[24] Gorkem Serbes, Sezer Ulukaya, and Yasemin Kahya. "An Automated Lung Sound Preprocessing and Classification System Based OnSpectral Analysis Methods". In: Jan. 2018, pp. 45–49.