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

Inception-Based Network and Multi-Spectrogram Ensemble Applied To Predict Respiratory Anomalies and Lung Diseases

Lam Pham[1], Huy Phan[2], Alexander Schindler[1], Ross King[1], Alfred Mertins[3], Ian McLoughlin[4]

Abstract— **This paper presents an inception-based deep neural network for detecting lung diseases using respiratory sound input. Recordings of respiratory sound collected from patients are first transformed into spectrograms where both spectral and temporal information are well represented, in a process referred to as front-end feature extraction. These spectrograms are then fed into the proposed network, in a process referred to as back-end classification, for detecting whether patients suffer from lung-related diseases. Our experiments, conducted over the ICBHI benchmark meta-dataset of respiratory sound, achieve competitive ICBHI scores of 0.53/0.45 and 0.87/0.85 regarding respiratory anomaly and disease detection, respectively.**

Clinical relevance— **Respiratory disease, wheeze, crackle, inception, convolutional neural network.**

I. INTRODUCTION

The World Health Organization has reported that one of the most common mortality factors worldwide is respiratory illness [1]. the most effective way to combat mortality from respiratory diseases is through early detection, which not only helps to limit the spread of infection but also improves treatment effectiveness. During a lung auscultation, which is an important aspect of a medical examination, experts can hear and detect anomalous sounds such as *Crackles* or *Wheezes* and thereby diagnose respiratory-relevant diseases. Therefore, if these anomaly sounds can be automatically detected by an edge device, it is very useful for self-observation, or early detection of such diseases. Analysing respiratory sound was mentioned in [2], [3], and recently this research topic has attracted considerable attention, with several machine learning methods having been proposed. In particular, frame-based systems proposed in [4] and [5] applied Mel-frequency cepstral coefficient (MFCC) extraction, a robust feature popularly used in Automatic Speech Recognition (ASR), to derive feature vectors. These vectors have been explored by conventional machine learning methods such as Hidden Markov Model [4], [5], Support Vector Machine [6], or Decision Tree [7]. Meanwhile, approaches relying on spectrogram representations involve generating two-dimensional spectrograms (i.e. an image), which are then fed into powerful network architectures such as CNN [8], [9] or RNN [10], [11] for classification. Although recent publications that have applied machine learning techniques report good performance, it is difficult to compare among systems due to the different training/test data ratios used as well as to the various experiments conducted over proprietary datasets. To make our work comparable, we evaluate our systems on the 2017 Internal Conference on Biomedical Health Informatics (ICBHI) [12], which is one of the largest public benchmark respiratory sound datasets. Furthermore, we obey the ICBHI challenge setup by using the ratio of 60/40 for training/test sets defined by the challenge [13], in which a subject is not presented in both training and test sets (note that some systems randomly separate ICBHI dataset into training and test subsets regardless of this patient interdependency [8], [9], [10], [14]). Regarding our proposed system, we firstly apply wavelet and gammatone transformations to generate a scalogram and a gammatonegram from an audio signal, respectively. These spectrograms are then fed into the proposed inception-based deep neural network to detect respiratory anomalies and lung diseases.

L. Pham, A. Schindler and R. King are with Competence Unit Data Science & Artificial Intelligence, Center for Digital Safety & Security, Austrian Institute of Technology, Austria.

H. Phan is with School of Electronic Engineering and Computer Science, Queen Mary University of London, UK.

A. Mertins is with the Institute for Signal Processing, University of Lübeck, Germany.

I. McLoughlin is with Singapore Institute of Technology, Singapore.

II. ICBHI DATASET AND TASKS DEFINED

The ICBHI dataset [12], which was collected from a total of 128 patients over 5.5 hours, comprises 920 audio recordings with a wide range of sampling frequencies ranging from 4 to 44.1 kHz and various lengths from 10 to 90 seconds (i.e. The challenge in [12] presents how the dataset was collected as well as recording devices used). In each recording, four different types of cycles (*Crackle*, *Wheeze*, *Both (Crackle & Wheeze)*, and *Normal*) are marked with onset and offset times. Additionally, each recording is also associated with the patient's disease status, mainly classified into three main categories: *Chronic Disease* (i.e. COPD, Bronchiectasis and Asthma), *Non-Chronic Disease* (i.e. Upper and Lower respiratory tract infection, Pneumonia, and Bronchiolitis), and *Healthy*. Given the ICBHI metadata, this paper proposes two main tasks. First, Task 1 aims to classify four different types of respiratory cycles mentioned. Second, Task 2, referred to as respiratory disease prediction, is to detect whether a patient suffers from *Chronic Diseases*, *Non-Chronic Diseases*, or *Healthy*. Regarding then experimental setting, we obey ICBHI challenge guidelines by splitting the audio recordings into training/test sets with a ratio of 60/40 without recordings of the same subject presenting in both training and test data (i.e. The challenge in [12] presents how to split the dataset into training/test subsets with non-overlapping patient subjects). While full audio recordings are evaluated in Task 2, respiratory cycles with onset and offset labels are extracted from all recordings for experiments conducted in Task 1. To evaluate performance and compare with state-of-the-art systems, we use metrics of Sensitivity (Sen.), Specificity (Spec.) and ICBHI scores (Average score (AS) and Harmonic score (HS)) that were comprehensively presented in [14], [15], [16].

III. Proposed baseline system

A. The baseline system architecture

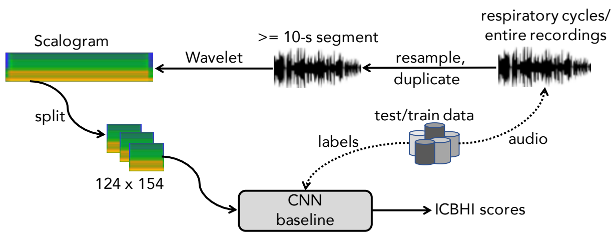

Fig. 1. The baseline system architecture.

TABLE I
THE CNN BASELINE NETWORK ARCHITECTURE PROPOSED

CNN baseline layers	Output ($W \times H \times C$)
BN - Conv [3×3] @ 64 - ReLU - BN - MP [2×2] - Dr (10%)	62×78×64
BN - Conv [3×3] @ 128 - ReLU - BN - MP [2×2] - Dr (15%)	31×39×128
BN - Conv [3×3] @ 256 - ReLU - BN - MP [2×2] - Dr (20%)	16×20×256
BN - Conv [3×3] @ 512 - ReLU - BN - GMP - Dr (25%)	512
FC - ReLU - Dr (30%)	1024
FC - Softmax	C

To evaluate the proposed system, we define a baseline as shown in Fig. 1 for comparison. In particular, respiratory cycles in Task 1 is re-sampled to 4 kHz as frequency bands of abnormal sounds (*Crackle* and *Wheeze*) located around 60-2000 Hz [10]. For the full recordings in Task 2, we re-sample them to 16 kHz to compensate for different recording sample rates. Re-sampled respiratory cycles or full recordings showing different lengths are next duplicated to ensure the same length of 10 seconds for respiratory cycles in Task 1 and minimum of 10 seconds in Task 2, respectively. Next, respiratory cycles go through a band-pass filter of 100-2000 Hz to reduce noise (note that band-pass filtering is not applied to the full recordings in Task 2). After that, these respiratory sounds are transformed into a scalogram by using continuous Wavelet transformation with *Morse* as the Wavelet mother function. Each 10-second scalogram of one respiratory cycle in Task 1 is thus scaled into an image of 124×154 image. Although the same scale ratio is also applied, the scalograms of full recordings in Task 2 show various time resolutions as the original recordings' lengths are different (note that the frequency resolution of 124 is identical for both tasks). Therefore, the long scalograms of full recordings in Task 2 are separated into various non-overlapped image patches of 124×154 that have the same size as 10-second scalograms in Task 1. To enlarge Fisher's criterion (i.e. the ratio of the between-class distance to the within-class variance in the feature space), we apply mixup data augmentation [17], [18] over image patches of 124×154 to increase variation of the training data.

For back-end classification, we propose a CNN-based network architecture shown in Table I, referred to as the CNN baseline. In particular, the CNN baseline contains sub-blocks which perform batch normalization (BN), convolution (Conv[kernel size] @ kernel number), rectified linear units (ReLU), max pooling (MP[kernel size]), global max pooling (GMP), dropout (Dr (percentage drop)), fully connected layers (FC), and Softmax configured as shown in the Table I. While the first FC layer is followed by ReLU and Dr, Softmax is used after the second FC layer to predict a probability among the categories classified. C takes values of 4 or 3 depending on the number of categories in Task 1 or Task 2, respectively.

B. Experimental setting for the baseline

As mixup data augmentation is used, labels are not represented in the one-hot encoding format. Therefore, we use Kullback–Leibler divergence (KL) loss shown in Eq. (1) below

$$Loss_{KL}(\Theta) = \sum_{n=1}^{N} \mathbf{y}_n \log\left(\frac{\mathbf{y}_n}{\hat{\mathbf{y}}_n}\right) + \frac{\lambda}{2}||\Theta||_2^2, \qquad (1)$$

where Θ are trainable parameters, constant λ is initially set to 0.0001, N is batch size set to 100, $\mathbf{y_i}$ and $\hat{\mathbf{y}}_\mathbf{i}$ denote expected and predicted results, respectively. We construct the proposed baseline with TensorFlow and the training is carried out for 100 epochs using Adam [19] for optimization.

As Task 2 evaluates over complete recordings while the proposed CNN baseline network works on one image patch of 124×154, the result over an entire recording is obtained by averaging the results over its patches. Let us consider $\mathbf{p}^m = (p_1^m, p_2^m, \ldots, p_C^m)$ as the probability obtained from the m^{th} out of M patches and let C be the number of categories classified. Then, the mean probability of a complete recording instance is denoted as $\bar{\mathbf{p}} = (\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_C)$ where

$$\bar{p}_c = \frac{1}{M}\sum_{m=1}^{M} p_c^m \quad \text{for} \quad 1 \le c \le C. \qquad (2)$$

The predicted label $\hat{y}$ is then determined as

$$\hat{y} = \underset{c \in \{1,2,\ldots,C\}}{\operatorname{argmax}} \; \bar{p}_c. \qquad (3)$$

IV. An analysis of inception-based network architecture and ensemble of multiple spectrogram input

Compared to the defined baseline, we evaluate whether the proposed inception-based network architecture and ensemble of different spectrograms are useful to improve the performance.

A. Inception-based deep neural network

Given the 10-second scalogram of a Wheeze cycle represented as an image with size of 124×154 as shown in Fig. 2 (note that the short-time Wheeze cycle is duplicated for three times to obtain 10-s duration in Fig. 2), the Wheeze spectrum is restricted within a narrow frequency band (i.e. the narrow

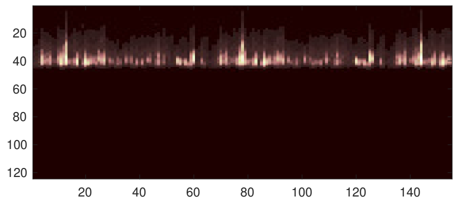

Fig. 2. 10-second scalogram of Wheeze cycles

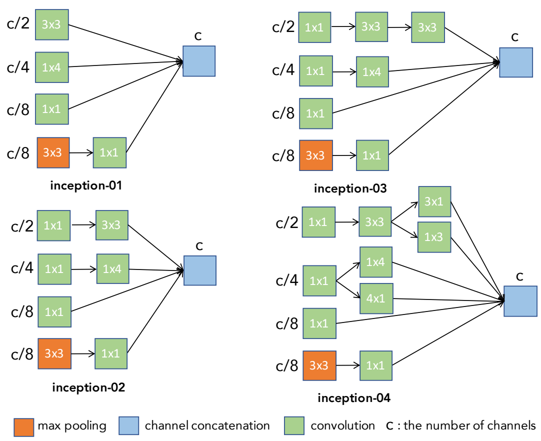

Fig. 3. Inception layer architectures.

band, a specific property of Wheeze sound, indices from 25 to 40 of 124 central frequencies distributed from minimum frequency of 100 Hz and maximum frequency of 2000 Hz) and shows short time duration (note that Crackle cycles also restricted to narrow frequency bands). This may cause ineffective if using a traditional CNN based network architecture with a single kernel size (i.e. the fixed kernel size of [3×3] is used popularly). To force the back-end classification model to learn these minor variations of spatial features in these narrow frequency bands, inception-based networks, which perform well on image data [20], are applied in this paper. In particular, we replace the convolutional layers (Conv) used in the CNN baseline by a different inception layer architectures as shown in Fig. 3. Notably, we use kernel [1×4] instead of [5×5] as usual to enforce the network focus on minor variation across the frequency dimension of the spectrum of Wheeze and Crackle sounds.

B. Ensemble of multiple spectrogram input

Inspired by [15], which shows that ensemble models of different spectrograms help to improve performance, we evaluate the combination of two scalogram (two Scal. for short) generated from two different Wavelet mother functions: *Morse* and *Amor*. While parameters in Morse function are set to obtain high resolution of frequencies, *Amor* function shows equal variance in time and frequency. We also evaluate another combination of scalogram (using *Morse* function) and gammatonegram (using gammatone filter [21]). Regarding the gammatone (Gam.) spectrogram, we use the setting of window size = 512, hop size = 256 and filter number = 124 to generate the same patch size of 124×154 as the scalogram (Scal.) mentioned in the baseline system in Section III. Meanwhile, the back-end classifier is reused from the baseline system proposed. To ensemble two baseline models, each of which learns from one type of input, we fuse the probabilities as in Eq. (4)

$$\bar{\mathbf{p}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{p}^k \tag{4}$$

where $\mathbf{p}^k$ is the probability output obtained from spectrogram k and $\bar{\mathbf{p}}$ is the probability output averaged over K spectrograms. Eventually, the final result is obtained by applying likelihood maximization in Eq. (3)

V. Experiments and Results

A. Effect of inception-based network architectures

TABLE II
EFFECT OF INCEPTION-BASED NETWORK ON RESPIRATORY ANOMALY DETECTION - TASK 1

Task	Systems	Spec.	Sen.	AS/HS Scores
Task 1	Baseline	0.68	0.30	0.49/0.42
Task 1	Inception-01	0.73	0.30	**0.52/0.43**
Task 1	Inception-02	0.70	0.30	0.50/0.42
Task 1	Inception-03	0.69	0.33	0.51/0.44
Task 1	Inception-04	0.70	0.32	0.51/0.44

TABLE III
EFFECT OF INCEPTION-BASED NETWORK ON RESPIRATORY DISEASE DETECTION - TASK 2

Task	Systems	Spec.	Sen.	AS/HS Scores
Task 2	Baseline	0.59	0.75	0.67/0.66
Task 2	Inception-01	0.88	0.81	0.85/0.84
Task 2	Inception-02	**1.00**	0.75	**0.87/0.85**
Task 2	Inception-03	0.53	**0.83**	0.68/0.64
Task 2	Inception-04	0.47	0.81	0.64/0.59

As shown in Tables II and III, the proposed inception-based network outperforms the CNN baseline for both classification tasks. In Task 1, the *inception-01* network achieves the best scores of 0.52/0.43. Meanwhile, the best scores of 0.87/0.85 are obtained by the *inception-02* architecture for Task 2 of lung disease detection (Note that the *inception-02* architecture can help to achieve the best Spec. score of 1.0).

B. Effect of multiple-spectrogram ensemble

TABLE IV
EFFECT OF MULTIPLE-SPECTROGRAM ENSEMBLE ON RESPIRATORY ANOMALY DETECTION - TASK 1

Task	Systems	Spec.	Sen.	AS/HS Scores
Task 1	Baseline	0.68	0.30	0.49/0.42
Task 1	Two Scal.	0.73	0.29	0.51/0.41
Task 1	Gam. & Scal.	0.72	0.31	**0.51/0.43**

TABLE V
EFFECT OF MULTIPLE-SPECTROGRAM ENSEMBLE ON RESPIRATORY DISEASE DETECTION - TASK 2

Task	Systems	Spec.	Sen.	AS/HS Scores
Task 2	Baseline	0.59	0.75	0.67/0.66
Task 2	Two Scal.	**0.65**	**0.79**	**0.72/0.71**
Task 2	Gam. & Scal.	**0.65**	0.76	0.70/0.70

Experimental results in Tables IV and V show that an ensemble of multiple spectrograms helps to improve the

TABLE VI

COMPARISON AGAINST STATE-OF-THE-ART SYSTEMS WITH ICBHI CHALLENGE SPLITTING - TASK 1 (HIGHEST SCORES IN **BOLD**).

Task	Method	Spec.	Sen.	AS/HS Scores
Task 1	DT [22]	0.75	0.12	0.43/0.15
Task 1	HMM [23]	0.38	**0.41**	0.39/0.23
Task 1	SVM [24]	0.78	0.20	0.47/0.24
Task 1	BRN [25]	0.69	0.31	0.50/0.43
Task 1	CNN-RNN [15]	**0.81**	0.28	**0.54/0.42**
Task 1	**Our system**	0.73	0.32	**0.53/0.45**

TABLE VII

COMPARISON AGAINST STATE-OF-THE-ART SYSTEMS WITH ICBHI CHALLENGE SPLITTING - TASK 2 (HIGHEST SCORES IN **BOLD**).

Task	Method	Spec.	Sen.	AS/HS Scores
Task 2	CRNN [26]	-	-	0.72/-
Task 2	CNN-MoE [16]	0.71	0.98	0.84/0.82
Task 2	**Our system**	0.88	0.85	**0.86/0.86**

performance compared to the baseline. While an ensemble of scalogram and grammatonegram achieves the best scores of 0.51/0.43 in Task 1, Task 2 shows the highest scores of 0.72/0.71 from an ensemble of two scalograms.

C. Performance Comparison to the state of the art

Given the results of inception-based networks and multiple spectrogram ensembles, we combine the *inception-01* network architecture and an ensemble of scalogram & gammatonegram for further analysis and compare the obtained results with the state-of-the-art systems (note that we only compare with systems that follow the standard ICBHI splitting of 60/40 with respect to subject independency [13]).

As shown by the comparison in Table VI, we achieve scores of 0.53/0.45 in Task 1 that are very competitive with the state-of-the-art systems. Task 2 results presented in Table VII show that the ensemble system outperforms the state-of-the-art systems, but is not better than the standalone system using *inception-02* (cf. Table III).

VI. CONCLUSION

This paper has presented an exploration of inception-based deep learning models and an ensemble of multiple input spectrograms for detecting respiratory anomaly and lung diseases from auditory recordings. By conducting extensive experiments on the ICBHI meta-dataset, we showed that our best model, which uses *inception-01* based architectures and ensemble of gammatonegram & scalogram, outperforms the state-of-the-art systems on both Task 1 and Task 2, thus validating the efficacy of deep learning for early diagnosis of respiratory diseases.

REFERENCES

[1] World Health Organization, "The global impact of respiratory diseases (second edition)," 2017. [Online]. Available: https://www.who.int/gard/publications/The_Global_Impact_of_Respiratory_Disease.pdf

[2] H. Polat and İ. Güler, "A simple computer-based measurement and analysis system of pulmonary auscultation sounds," *Journal of medical systems*, vol. 28, no. 6, pp. 665–672, 2004.

[3] R. J. Riella, P. Nohama, R. F. Borges, and A. L. Stelle, "Automatic wheezing recognition in recorded lung sounds," in *Proc. EMBC*, 2003, pp. 2535–2538.

[4] H. Yamamoto, S. Matsunaga, M. Yamashita, K. Yamauchi, and S. Miyahara, "Classification between normal and abnormal respiratory sounds based on stochastic approach," in *Proc. 20th International Congress on Acoustics*, 2010, pp. 4144–4148.

[5] T. Okubo, N. Nakamura *et al.*, "Classification of healthy subjects and patients with pulmonary emphysema using continuous respiratory sounds," in *Proc. EMBC*, 2014, pp. 70–73.

[6] M. Grønnesby, J. C. A. Solis, E. Holsbø, H. Melbye, and L. A. Bongo, "Feature extraction for machine learning based crackle detection in lung sounds from a health survey," *arXiv preprint arXiv:1706.00005*, 2017.

[7] G. Chambres, P. Hanna, and M. Desainte-Catherine, "Automatic detection of patient with respiratory diseases using lung sound analysis," in *Proc. CBMI*, 2018, pp. 1–6.

[8] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 65, 2017.

[9] D. Perna, "Convolutional neural networks learning from respiratory data," in *Proc. BIBM*, 2018, pp. 2109–2113.

[10] D. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *Proc. CBMS*, 2019, pp. 50–55.

[11] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *International Conference on Artificial Neural Networks*, 2018, pp. 208–217.

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

[13] I. challenge, *Dataset splitting ratio*, https://bhichallenge.med.auth.gr/sites/default/files/ICBHI_final_database/ICBHI_challenge_train_test.txt.

[14] L. Pham, I. McLoughlin, H. Phan, M. Tran, T. Nguyen, and R. Palaniappan, "Robust deep learning framework for predicting respiratory anomalies and diseases," in *Proc. EMBC*, 2020, pp. 164–167.

[15] K. Minami, H. Lu, H. Kim, S. Mabu, Y. Hirano, and S. Kido, "Automatic classification of large-scale respiratory sound dataset based on convolutional neural network," in *Proc. ICCAS*, 2019, pp. 804–807.

[16] L. Pham, H. Phan, R. Palaniappan, A. Mertins, and I. Mcloughlin, "Cnn-moe based framework for classification of respiratory anomalies and lung disease detection." *IEEE journal of biomedical and health informatics*, 2021.

[17] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acoustic scene classification using multi-channel convolutional neural network," in *Pacific Rim Conference on Multimedia*, 2018, pp. 14–23.

[18] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," in *ICLR*, 2018.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[21] D. Ellis, "Gammatone-like spectrogram," 2009. [Online]. Available: http://www.ee.columbia.edu/dpwe/resources/matlab/gammatonegram

[22] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, vol. 40, no. 3, p. 035001, 2019.

[23] N. Jakovljević and T. Lončar-Turukalo, "Hidden markov model based respiratory sound classification," in *Precision Medicine Powered by pHealth and Connected Health*. Springer, 2018, pp. 39–43.

[24] G. Serbes, S. Ulukaya, and Y. P. Kahya, "An automated lung sound preprocessing and classification system based on spectral analysis methods," in *Precision Medicine Powered by pHealth and Connected Health*, 2018, pp. 45–49.

[25] Y. Ma, X. Xu, Q. Yu, Y. Zhang, Y. Li, J. Zhao, and G. Wang, "Lungbrn: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2019, pp. 1–4.

[26] 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.