2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871440

An Ensemble of Deep Learning Frameworks for Predicting Respiratory Anomalies

Lam Pham^{1*}, Dat Ngo^{2*}, Khoa Tran^{3}, Truong Hoang^{4}, Alexander Schindler^{1}, Ian McLoughlin^{5}

Abstract— **This paper evaluates a range of deep learning frameworks for detecting respiratory anomalies from input audio. Audio recordings of respiratory cycles collected from patients are transformed into time-frequency spectrograms to serve as front-end two-dimensional features. Cropped spectrogram segments are then used to train a range of back-end deep learning networks to classify respiratory cycles into predefined medically-relevant categories. A set of those trained high-performance deep learning frameworks are then fused to obtain the best score. Our experiments on the ICBHI benchmark dataset achieve the highest ICBHI score to date of 57.3%. This is derived from a late fusion of inception based and transfer learning based deep learning frameworks, easily outperforming other state-of-the-art systems.**

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

I. INTRODUCTION

Automated respiratory sound analysis (ARSA) has recently attracted much research attention, encouraged by advances in robust machine and deep learning technologies, which can be leveraged into this important application area. Systems proposed by authors generally comprise two main steps, referred to as front-end feature extraction and back-end modelling. In machine learning based systems, hand-crafted features such as Mel-frequency cepstral coefficients (MFCC) [1], [2], or a combination of several time domain features (e.g. variance, range, sum of simple moving average) and frequency domain features (e.g. spectrum mean) [3] are extracted during the front-end feature extraction. These features are then fed into conventional machine learning models, such as Hidden Markov Models [2], Support Vector Machines [3], or Decision Trees [1] for specific tasks of classification or regression. Meanwhile, deep learning based systems make use of raw inputs such as waveforms or spectrograms, with a trained feature extractor. Spectrograms, in which both temporal and spectral feature elements are well represented, have been explored by a wide range of deep and convolutional neural networks (CNNs) [4], [5], [6], [7] and recurrent neural networks (RNNs) [8]. Comparing between machine learning approaches with hand-crafted features, and deep learning systems with trained feature extractors, the latter are widely reported as being more effective for respiratory classification tasks [4], [6], [7].

We therefore evaluate a wide range of deep leaning frameworks with spectrogram inputs, trained for the specific task of audio respiratory cycles classification, and then their fusion at three levels. We conduct extensive experiments using the 2017 ICBHI (International Conference on Biomedical Health Informatics) dataset [9], which is one of the largest benchmark respiratory sound datasets and widely used in comparative studies. Our main contributions are (1) We evaluate whether benchmark and complex deep neural network architectures (e.g. ResNet50, Xception, InceptionV3, etc.) are more effective than inception based and low footprint models, and (2) We evaluate whether applying transfer learning techniques on the downstream task of respiratory cycle classification can achieve competitive performance over direct training approaches.

II. ICBHI DATASET AND TASKS DEFINED

The ICBHI dataset [9] is comprised of 920 separate audio recordings collected from 128 patients over 5.5 hours. Each audio recording contains one or different types of respiratory cycles, labeled as *Crackle*, *Wheeze*, *Both Crackle & Wheeze*, or *Normal*. The labels, determined by respiratory experts, have fine resolution onset and offset times. The dataset is considered relatively well-labeled, and since recordings are made by a variety of instruments, and are sometimes acoustically noisy, it is reflective of real-world conditions. Given this ICBHI dataset, the current paper aims to classify the four different types of respiratory cycles mentioned – and that is also the main task of the ICBHI challenge itself [9]. To evaluate, we adhere to the ICBHI challenge settings, splitting audio recordings into Train and Test subsets with a ratio of 60/40 without overlapping patient in both subsets (please note that some published systems randomly separate the ICBHI recordings into training and test subsets regardless of the source patient, so on those systems, recordings from the same patient can occur in both training and test sets [4], [5]. By contrast, we ensure no patient overlap between sets). Using reported onset and offset times, we then extract respiratory cycles from entire recordings, to obtain four categories of respiratory cycles on each subset. Regarding the evaluation metrics, we use Sensitivity (Sen.), Specitivity (Spec.), and ICBHI score (ICB.) which is the mean of the Sen. and Spec. scores. These scores are the same as those required by the ICBHI challenge [10] and [11], [12].

L. Pham and A. Schindler are with Center for Digital Safety & Security, Austrian Institute of Technology, Austria.

D. Ngo is with School of Computer Science and Electronic Engineering, University of Essex, UK.

K. Tran is with Faculty of electrical engineering, University of Science and Technology, the University of Danang, Viet Nam.

T. Hoang is with AI Center, FPT Software Company Limited, Vietnam.

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

(*) Main and equal contribution into the paper.

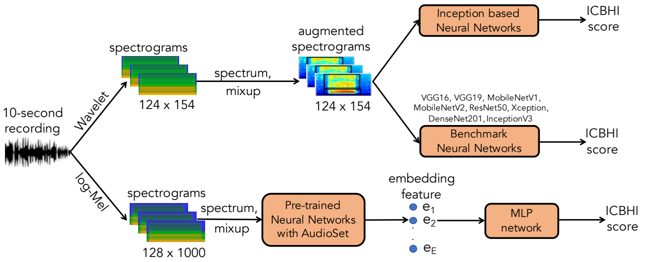

Fig. 1. High-level architecture of three proposed deep learning frameworks.

TABLE I

THE GENERAL INCEPTION BASED NETWORK ARCHITECTURES.

Single Inception Layer	**Double Inception Layers**
BN	
Inc(**ch1**) - ReLU	Inc(**ch1**) - ReLU - Inc(**ch1**) - ReLU
BN - MP - Dr(10%) - BN	
Inc(**ch2**) - ReLU	Inc(**ch2**) - ReLU - Inc(**ch2**) - ReLU
BN - MP - Dr(15%) - BN	
Inc(**ch3**) - ReLU	Inc(**ch3**) - ReLU - Inc(**ch3**) - ReLU
BN - MP - Dr(20%) - BN	
Inc(**ch4**) - ReLU	Inc(**ch4**) - ReLU - Inc(**ch4**) - ReLU
BN - GMP - Dr(25%)	
FC(**fc1**) - ReLU - Dr(30%)	
FC(**fc2**) - ReLU - Dr(30%)	
FC(**4**) - Softmax	

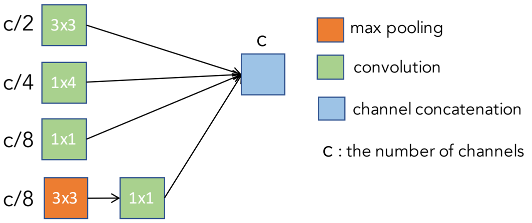

Fig. 2. The single inception layer architecture.

TABLE II

SETTING FOR INCEPTION BASED NETWORK ARCHITECTURES.

Networks	**Inc-01**	**Inc-02**	**Inc-03**	**Inc-04**	**Inc-05**	**Inc-06**
Single/Double	Single	Double	Single	Double	Single	Double
ch1	32	32	64	64	128	128
ch2	64	64	128	128	256	256
ch3	128	128	256	256	512	512
ch4	256	256	512	512	1024	1024
fc1	512	512	1024	1024	2048	2048
fc2	512	512	1024	1024	2048	2048

III. DEEP LEARNING FRAMEWORKS PROPOSED

To classify four types of respiratory cycles from the ICBHI dataset, we firstly propose a high-level architecture of three main deep learning frameworks as shown in Fig. 1, which contain the following:

I The upper stream in Fig. 1 shows how we directly train small-footprint inception based network architectures from augmented spectrograms.

II Benchmark and large footprint deep learning network architectures of VGG16, VGG19, MobileNetV1, MobileNetV2, ResNet50, DenseNet201, InceptionV3, Xception are directly trained and evaluated as shown in the middle stream of Fig. 1.

III The lower stream in Fig. 1 shows how we reuse pre-trained models, which were trained with the large-scale AudioSet to extract embedding features. These are used in turn to train a multilayer perceptron (MLP) network for the final classification.

In general and as mentioned previously, these three deep learning frameworks each comprise two main steps of front-end spectrogram-derived feature extraction, followed by a back-end classification model.

A. The front-end spectrogram feature extraction

The input to the proposed deep learning frameworks shown in Fig. 1 are 10 second recorded segments of respiratory cycles. During training, since cycles naturally have a range of durations, we duplicate shorter cycles or truncate longer cycles to provide equal-dimension audio input segments. For the first two deep learning frameworks (I) and (II), we extract Wavelet-based spectrograms, which had proven effective in our previous work [7], and reuse the same optimal extraction settings from [7]. We then generate Wavelet spectrograms of dimension 124×154 from each 10 second respiratory cycle segment. For the deep learning framework (III), we extract log-Mel spectrograms since we employ pre-trained models from [13] which require a log-Mel spectrogram input. By using the same settings, we generate log-Mel spectrograms of dimension 128×1000 from each 10 second respiratory cycle segment. To improve the back-end classifier performance, two data augmentation methods are employed for all frameworks. Specifically, spectrum [14] and mixup [15] augmentation are applied on both log-Mel and Wavelet-based spectrogram inputs before feeding into the back-end deep learning models for classifier training.

B. The back-end deep learning classifier networks

(I) The low-footprint inception based network architectures: Since good results were achieved using an inception based network in our previous work [7], we further evaluate different types of inception based network architectures in this paper. In particular, two high-level architectures with single or double inception layers are explored, as defined in Table I. These architectures comprise of several different layer types. The inception layer (Inc(output channel number)) is shown in Fig. 2, and also includes batch normalization (BN), rectified linear units (ReLU), max pooling (MP), global max pooling (GMP), dropout (Dr(percentage)), fully connected (FC(output node number)) and Softmax layer types. By using the two architectures and setting adjusting parameters such as channel numbers of inception layers and output node numbers of fully connected layers, we create six inception based deep neural network variants as shown in Table II, referred to as Inc-01 to Inc-06, respectively.

(II) The benchmark and complex neural network architectures: We next evaluate different benchmark neural network architectures, namely VGG16, VGG19, MobileNetV1, MobileNetV2, ResNet50, DenseNet201, InceptionV3, and Xception, which are available in the Keras library [16] and are popularly applied in different research domains. Compared with the inception based network architectures used in framework (I), these benchmark neural networks have a larger footprint and a more complex architecture consisting of trunks of convolutional layers.

TABLE III

PERFORMANCE COMPARISON OF PROPOSED DEEP LEARNING FRAMEWORKS.

Inception based Frameworks	Scores (Spec./Sen./ICB.)	Benchmark Frameworks	Scores (Spec./Sen./ICB.)	Transfer learning Frameworks	Scores (Spec./Sen./ICB.)
Inc-01	56.3/**40.5**/48.4	VGG16	70.1/28.6/49.3	VGG14	**82.1**/28.1/**55.1**
Inc-02	69.7/31.9/50.8	VGG19	69.7/28.4/49.1	DaiNet19	76.4/26.9/51.7
Inc-03	81.7/28.4/**55.1**	MobileNetV1	75.5/14.3/44.9	MobileNetV1	64.4/**40.3**/52.3
Inc-04	**84.0**/24.8/54.4	MobileNetV2	74.7/16.1/45.4	MobileNetV2	76.0/32.7/54.4
Inc-05	80.5/26.3/53.4	ResNet50	**88.0**/15.2/**51.6**	LeeNet24	70.7/30.9/52.8
Inc-06	74.8/30.0/52.4	DenseNet201	71.7/30.3/51.1	Res1DNet30	74.9/26.7/50.8
		InceptionV3	70.9/**32.2/51.6**	ResNet38	71.6/32.2/51.9
		Xception	75.7/22.1/48.9	Wavegram-CNN	69.0/38.1/53.5

TABLE IV

THE MLP ARCHITECTURE USED FOR TRAINING EMBEDDING FEATURES.

Setting layers	Output
FC(4096) - ReLU - Dr(10%)	4096
FC(4096) - ReLU - Dr(10%)	4096
FC(1024) - ReLU - Dr(10%)	1024
FC(4) - Softmax	4

(III) The transfer learning based network architectures: As transfer learning techniques have proven effective for downstream tasks with a limitation of training data and smaller categories classified [13], we leverage pre-trained networks which were trained with the large-scale AudioSet dataset from [13]: LeeNet24, DaiNet19, VGG14, MobileNetV1, MobileNetV2, Res1DNet30, ResNet38, Wavegram-CNN. We then modify these networks to match the downstream task of classifying the four respiratory cycles of the ICBHI dataset.

In particular, we retain trainable parameters from the first layer to the global pooling layer of the pre-trained networks. We then replace layers after the global pooling layer by new fully connected layers to create a new network (i.e. the trainable parameters in new fully connected layers are initialized with random values of mean 0 and variance 0.1). In the other words, we use a multilayer perceptron (MLP) as shown in Table IV. This contains FC, ReLU, Dr, and Softmax layers and is trained using embedding features extracted from the pre-trained models. Hence, the embedding features are the feature map of the final global pooling layer in the pre-trained network.

IV. EXPERIMENTS AND RESULTS

A. Experimental setting for back-end classifiers

Due to use of the spectrum [14] and mixup [15] data augmentation methods, labels are no longer in one-hot encoding format. Therefore, we use a 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 set initially to 0.0001, batch size N is set to 100, $\mathbf{y_i}$ and $\hat{\mathbf{y}}_\mathbf{i}$ denote expected and predicted results. While we construct deep learning networks proposed in frameworks (I) and (II) with Tensorflow, we use Pytorch for extracting embedding features and training MLP in the framework (III), since the pre-trained networks used were built in a Pytorch environment. We use the Adam method for optimization and train for 100 epochs.

TABLE V

PERFORMANCE COMPARISON OF FUSION STRATEGIES OF INCEPTION BASED AND TRANSFER LEARNING BASED FRAMEWORKS.

Fusion strategies	Spec.	Sen.	ICB.
Pre-trained VGG14	82.1	28.1	55.1
Inc-03	81.7	28.4	55.1
The early fusion	79.9	**30.9**	55.4
The middle fusion	**87.3**	25.1	56.2
The late fusion	85.6	29.0	**57.3**

B. Performance comparison among deep learning frameworks proposed

From the experimental results shown in Table III, it can be seen that generally the low-footprint inception based frameworks and the transfer learning based frameworks are competitive and outperform the benchmark frameworks. Table III records the best ICBHI score of 55.1% from the Inc-03 framework, matched by the transfer learning framework using a pre-trained VGG14. The best performance obtained from the pre-trained VGG14 makes sense as this network outperforms the other network architectures for classifying sound events in the AudioSet dataset. Notably, while we use the same network architecture of MobileNetV1 and MobinetV2 for both benchmark and transfer learning framework, we see that the latter significantly outperforms the former. From these results we can conclude that (1) applying the transfer learning technique on the downstream task of classifying respiratory cycles is effective; and (2) low-footprint Inception based networks, focusing on minimal variation of time and frequency, are effective for respiratory sounds – outperforming the larger and more complex benchmark architectures.

C. Early, middle, and late fusion of inception based and transfer learning based frameworks

As the deep learning frameworks basing on Inc-03 and transfer learning with the pre-trained VGG14 achieve the best scores, we then evaluate whether a fusion of results from these frameworks can help to further improve the task performance. In particular, we propose three fusion strategies to compare. In the first and second fusion strategies, referred to as the early and middle fusion, we concatenate the embedding feature extracted from the pre-trained VGG14 (e.g. the feature map of the global pooling of the pre-trained VGG14) with the embedding feature extracted from Inc-03 to generate a new combined feature. We then train the new combined feature with an MLP network architecture, as shown in Table IV. While the feature map of the max global pooling (MGP) of Inc-03 is considered as the embedding feature in the first fusion strategy, the feature map of the

TABLE VI

COMPARISON AGAINST STATE-OF-THE-ART SYSTEMS.

Method	Spec.	Sen.	ICBHI Score
HMM [18]	38.0	41.0	39.0
DT [9]	75.0	12.0	43.0
1D-CNN [17]	36.0	51.0	43.0
SVM [19]	78.0	20.0	47.0
Autoencoder [20]	69.0	30.0	49.0
ResNet [21]	63.2	41.3	52.3
Inception [7]	73.2	32.2	53.2
CNN-RNN [11]	81.0	28.0	54.0
ResNet50 [22]	72.3	40.1	56.2
Our best system	**85.6**	**29.0**	**57.3**

second fully connected layer of Inc-03 (e.g. FC(fc2)) is used by the second fusion strategy. In the third fusion strategy, referred to as the late fusion, we use a product fusion of the predicted probabilities obtained from these inception based and transfer learning based frameworks. The product fusion result $\mathbf{p_{f-prod}} = (\bar{p}_1, \bar{p}_2, ..., \bar{p}_C)$ is obtained by:

$$\bar{p_c} = \frac{1}{S}\prod_{s=1}^{S}\bar{p}_{sc} \quad for \quad 1 \leq s \leq S, \tag{2}$$

where $\mathbf{\bar{p}_s} = (\bar{p}_{s1}, \bar{p}_{s2}, ..., \bar{p}_{sC})$ is the predicted probability of a single framework, C is the category number and the s^{th} out of S individual frameworks evaluated. The predicted label $\hat{y}$ is determined by:

$$\hat{y} = argmax(\bar{p}_1, \bar{p}_2, ..., \bar{p}_C) \tag{3}$$

Results are compared in Table V, which shows that all three fusion strategies can enhance performance, improving ICBHI score by 0.3, 1.1, 2.2 for early, middle and late fusion respectively. This indicates that embedding features extracted from the Inception based Inc-03 framework and from the transfer learning framework with the pre-trained VGG14, contain distinct and somewhat complimentary features to describe respiratory cycles.

D. Performance comparison to the state of the art

To compare against the state of the art, we only select published systems which follow the recommended setting of the ICBHI challenge [10] with a train/set ratio of 60/40 and no overlapping patient subjects between the two subsets. These experimental results are shown in Table VI. It can be seen that the final proposed system using a late fusion of inception based and transfer learning frameworks outperforms the state of the art, recording the best score of 57.3%. However, the best Sen. score from [17] reports 51.0, which shows the ICBHI dataset challenging and requires further research for enhancing the performance.

V. CONCLUSION

This paper has presented an exploration of various deep learning models for detecting respiratory anomalies from auditory recordings. We consider three frameworks of deep learning for this task, encompassing a very wide range of different networks and architectures, and consider their fusion, obtained at three different levels. We conduced extensive experiments using the ICBHI dataset (operating with ICBHI challenge settings), to compare between networks and settings. Eventually, we found that our best proposed model uses a late product-based fusion of Inception-derived and transfer learning frameworks. The resulting ICBHI score easily outperforms state-of-the-art published systems, including many benchmark frameworks, thus validating this application of deep learning for the detection of respiratory anomalies.

REFERENCES

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

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

[3] 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," *preprint arXiv:1706.00005*, 2017.

[4] 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, pp. 65, 2017.

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

[6] 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.

[7] L. Pham, H. Phan, R. King, A. Mertins, and I. McLoughlin, "Inception-based network and multi-spectrogram ensemble applied for predicting respiratory anomalies and lung diseases," in *Proc. EMBC*, 2021, pp. 253–256.

[8] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *Proc. ICANN*, 2018, pp. 208–217.

[9] 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. 2018.

[10] ICBHI 2017 Challenge, `https://bhichallenge.med.auth.gr/sites/default/files/ICBHI_final_database/ICBHI_challenge_train_test.txt`.

[11] 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.

[12] L. D. 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*, vol. 25, no. 8, pp. 2938–2947, 2021.

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

[14] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *preprint arXiv:1904.08779*, 2019.

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

[16] F. Chollet et al., "Keras library," `https://keras.io`, 2015.

[17] P. Faustino, J. Oliveira, and M. Coimbra, "Crackle and wheeze detection in lung sound signals using convolutional neural networks," in *Proc. EMBC*, 2021, pp. 345–348.

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

[19] 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*, pp. 45–49. Springer, 2018.

[20] D. Ngo, L. Pham, A. Nguyen, B. Phan, K. Tran, and T. Nguyen, "Deep learning framework applied for predicting anomaly of respiratory sounds," in *Proc. ISEE*, 2021, pp. 42–47.

[21] Y. Ma, X. Xu, and Y. Li, "LungRN+NL: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation.," in *Proc. INTERSPEECH*, 2020, pp. 2902–2906.

[22] S. Gairola, F. Tom, N. Kwatra, and M. Jain, "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," in *Proc. EMBC*, 2021, pp. 527–530.