A Novel Melspectrogram Snippet Representation Learning Framework For Severity Detection of Chronic Obstructive Pulmonary Diseases

SUBMISSION DATE / POSTED DATE

20-12-2022 / 24-12-2022

CITATION

Roy, Arka; Satija, Udit (2022): A Novel Melspectrogram Snippet Representation Learning Framework For Severity Detection of Chronic Obstructive Pulmonary Diseases. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.21758660.v1

DOI

10.36227/techrxiv.21758660.v1

A Novel Melspectrogram Snippet Representation Learning Framework For Severity Detection of Chronic Obstructive Pulmonary Diseases

Arka Roy, Udit Satija, *Senior Member, IEEE*

***Abstract*—Chronic obstructive pulmonary disease (COPD) is a major public health concern across the world. Since it is an incurable disease, early detection and accurate diagnosis are very crucial for preventing the progression of the disease. Lung sounds provide reliable and accurate prognoses for identifying respiratory diseases. Recently, Altan et al. recorded 12-channel real-time lung sound dataset, namely RespiratoryDatabase@TR, for five different severity levels of COPD at Antakya State Hospital Turkey, and proposed deep learning frameworks for two-class COPD classification and five-class classification using a deep belief network (DBN) classifier and extreme learning machine (ELM) classifier respectively. A classification accuracy of 95.84% and 94.31% were achieved for two-class and five-class respectively. In this paper, we have proposed a melspectrogram snippet representation learning framework for both two-class and five-class COPD classification. The proposed framework consists of the following stages: preprocessing, melspectrogram snippet representation generation from lung sound and fine tuning of a pretrained YAMNet. Experimental analysis on the RespiratoryDatabase@TR dataset demonstrates that the proposed framework achieves accuracies of 99.25% and 96.14% for binary and multi-class COPD severity classification respectively, which is superior to the only existing methods proposed by Altan et al. for severity analysis of COPD using lung sounds.**

***Index Terms*—Chronic Obstructive Pulmonary Disease, Lung Sounds, YAMNet, Transfer Learning.**

I. Introduction

RESPIRATORY diseases are the world's third leading cause of mortality. Each year, more than 3 million individuals worldwide die as a result of one of the five major pulmonary disorders [1] (asthma, COPD, tuberculosis, lung cancer, and lower respiratory tract infection) [2]. In particular, COPD is a major public health concern across the world as it is incurable and takes a considerably long time to get diagnosed. For an experienced pulmonologist also, it is critical to provide straightforward analytical anomalies. In general, the lungs' airways and alveoli are elastic or flexible. However, in the case of COPD, less air passes in and out of the airways because: (a) the airway walls are thickened and become inflammatory; and (b) the airways create more mucus than usual and might become blocked. COPD can be characterized by adventitious breathing which can be observed during lung auscultation. Due to the severe consequences of COPD, early detection and accurate diagnosis is very crucial. The majority of the prior research stand out because these works incorporate several pathological variables, including spirometry measures, age, sex, blood pressure, heart rate, hemoglobin, hematocrit, etc. for COPD severity classification [3], [4], [5].

Spirometry, often known as lung function test is the gold standard technique for COPD diagnosis [6]. Spirometry evaluates how quickly and how much air a person can exhale. A patient who is undergoing a spirometry test must inhale deeply and exhale as quickly as forcefully as they can into a mouthpiece while having their nose clipped. Recently, at-home spirometry has become prevalent for regular monitoring of lung capacity with the emergence of over-the-counter tabletop to portable spirometers [7]. To diagnose and assess the severity of COPD, values for spirometry variables including forced expiratory volume in 1 second (FEV1), forced expiratory capacity (FVC), and FEV1/FVC ratio or forced expiratory volume ratio (FEVR) are employed [6]. The most common symptom of COPD during lung auscultation is wheezing, which is caused by narrowed airways and blockages caused by sputum (mixture of saliva and mucus). It's a chronic condition characterized by musical clearing on both exhale and inhalation [8]. Normal or vesicular respiratory sounds are mild and perceptible in both inspiratory and expiratory phase with a frequency range of 100 to 1kHz. However, wheezes are loud, high-pitched, and have a frequency range of more than 400 Hz [8]. Based on the spirometric measurement: FEVR and the wheezing characteristics from the respiratory sound, the global initiative for chronic obstructive lung disease (GOLD) has divided the degree of COPD severity into 5 groups [9]: COPD0, COPD1, COPD2, COPD3, COPD4. COPD0 is a low-risk condition for those who have been smoking for few years, having FEVR more than 85 percent [9]. Patients with moderate level COPD (COPD1) exhibit prolonged symptoms with minimal wheezing, and they have FEVR of more than 80% [9]. Whereas, COPD2 refers to those who have a intermediate degree of COPD severity, with FEVR ranging from 50% to 80% [9]. Significant wheezing during expiration and inspiration is typical in COPD3 and COPD4. It is caused by blockage and constriction of the airways, as well as usually coexisting heart disorders. People with a FEVR of 30% to 50% with all chronic symptoms, as well as lung infections, are likely to have severe COPD and fall into the COPD3 group [9]. Patients with COPD4, the most severe stage of COPD, have a FEVR of less than 30% and experience all of the chronic symptoms and respiratory issues [9]. However, spirometry is dependent on patient efforts, cooperation with the technician, as it is a highly laborious procedure [10], [7].

Arka Roy and Udit Satija are with the Department of Electrical Engineering, Indian Institute of Technology Patna, Bihar-801106, India (e-mail: arka_2121ee34@iitp.ac.in, udit@iitp.ac.in).

A. Related work and motivation

Lung auscultation is one of the most popular and traditional method used by the pulmonary specialists to analyze the status of the respiratory system [11]. Although doctors use photoplethesmogram [12], spirometry [6] etc., lung auscultation remains essential to doctors due to its simple and cost effective nature [11]. Lung sounds are connected to anatomical flaws in the lungs, making lung sound examinations more trustworthy and accurate for detecting respiratory disorders. Various pathological lung sounds provide different clues to the experts regarding a variety of pulmonary disorders. For instance COPD is identified by wheezing during auscultation. As per GOLD standard, spirometry is used for COPD detection [10]. However, spirometry values are dependent on patient efforts, cooperation with the technician, and interest as it is a highly laborious procedure, especially for younger and the elderly [10]. Thereby, developing AI-based automated algorithms using lung sound signals will be extremely beneficial in detection, staging of COPD and will be suitable for people of all ages. However, very few research works have been carried out on COPD severity detection from lung sounds, as majority of research have focused on identifying adventitious sound anomalies [13], [14], [15] , [16] rather than diagnosing chronic respiratory diseases by utilizing the lung sounds [17], [18], [19]. On the other hand, most of the COPD categorization works focus on clinical data [3], [4], [5], [20] which are heavily dependent on clinical technician [10]. This highlights the motivation for including and intervening with lung sounds in COPD severity grading.

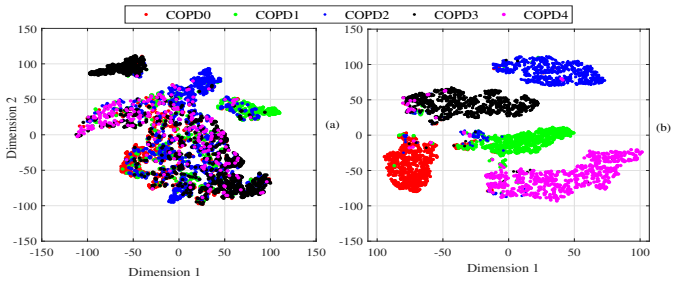

Fig. 1. Two dimensional visualisation of (a) Original lung sound signals, and (b) Lung sounds after employing proposed framework using t-SNE.

Naves et al. [21] used high-order statistical characteristics (cumulants) to examine diseased lung sounds. To increase classification performance, linear as well as nonlinear classification algorithms were applied to identify crackle and wheeze lung sounds. Fernandez-Granero et al. [22] examined tracheal sounds for early detection of COPD. They had achieved highest classification accuracy of 75.80% by employing a neural network architecture. Oweis et al. [23] performed an adventitious lung sound classification task by feeding the power spectral density derived features along with morphological features to a neural network. Newandee et al. [20] suggested a technique for diagnosing COPD based on heart rate variability (HRV) readings using principal component analysis (PCA) and clustering methods. An accuracy of 88% was achieved in classifying COPD and non-COPD individuals. On the basis of clinical data, Amaral et al. [24] suggested a COPD diagnosis framework with 90% classification accuracy. To identify the different distinctive information present in respiratory sounds, Morillo et al. [25] investigated the time-frequency representation derived from the short-time Fourier transform (STFT). Following that, COPD was detected with an accuracy of 81.80% by utilizing an artificial neural network (ANN) based classification model. In recent years, deep learning techniques have also been employed in COPD severity classification [26], [27], [17], [18], [19]. Sugimori et al. [28] had used CT scan images to classify five class COPD severities using convolutional neural network (CNN). They have used a ResNet50 [29] architecture as CNN backbone follwed by dense classifier and achieved a classification accuracy of 44% for five class COPD severity classification. A new non-invasive imaging approach, named parametric response mapping (PRM) [30], [26] is employed in conjunction with inspiratory and expiratory CT images for early detection of small airway damages occurred due to COPD. Ho et al. [26] have used 3D-CNN architecture in combination with PRM to classify subjects affected with COPD. They have used two functional parenchyma variables named: emphysema percentage and small airway disease percentage as inputs of 3D-CNN architecture and received an accuracy rate of 89.3% and sensitivity of 88.3% respectively, while classifying COPD subjects from non-COPD ones. Altan et al. [17] have classified COPD and healthy patients by using 12-channel lung sound based on ensemble empirical mode decomposition and extracted statistical as well as time domain features. An accuracy rate of 93.67% is achieved using DBN classifier. Altan et al. [18] proposed a 3D second order difference plot (SODP) based feature extraction strategy that use a 3D second order difference plot methodology and DBN classifier to distinguish between two extreme severity levels: COPD0 and COPD4. Using these features, 95.84% accuracy was achieved on two class severity classification. However, these two severities can be easily differentiated by analyzing the wheezing density. In another work, Altan et al. [19] employed 3D-SODP to extract distinctive anomalies from the lung sound data, and used cuboid , octant quantisation of the 3D-SODP features. Thereafter, these features were fed to a deep ELM model to categorize all five severity levels of COPD. The deepELM architecture uses lower-upper triangular ELM (LuELM) and Hessenberg ELM kernels for achieving better generalization and faster training convergence. By employing this deep ELM architecture, a classification accuracy of 94.31%, a weighted sensitivity of 94.28%, and a weighted specificity of 98.76% was achieved for five class COPD severity classification task.

B. Objective and key contributions

The primary objective of the study is to propose a deep learning framework based on lung sounds, that can classify all five severity levels of COPD, which are hard to differentiate without performing additional tests like the spirometry test. Early detection of COPD is critical for preventing disease development and improving people's quality of life. However, distinguishing the following severities: COPD0, 1 and

2, without employing any further diagnostic tests is practically impossible. The same problem persists in identifying COPD3 and 4 as both of these classes have almost similar properties. Due to the similar characteristics of lung sounds, even an experienced pulmonologist specialist may make a mistake during an auscultation examination if they do not use other intelligent diagnostic techniques. To visualize the problem, t-distributed stochastic neighbor embedding (t-SNE) plot is used, which depicts lung sound data of different classes of COPD in a two-dimensional space [31]. In Fig. 1(a), it can be observed that initially all the lung sounds belonging to different classes are not distributed in separate clusters. Therefore, there is a need to develop an intelligent deep learning framework that can distinguish different COPD severities. The proposed framework exploits the potential of the fine-tuning process of a pretrained YAMNet [32] for COPD severity categorization, which is trained on mel-spectrograms of audio signals present in AudioSet [33], the largest dataset for audio machine learning. The proposed framework involves the following stages: (a) preprocessing of the lung sound signals; and (b) melspectrogram snippet representation learning based severity classification, which contains two sub-modules: 1) where, the processed lung sound signal is converted to vanilla melspectrogram and then framed into sub-images, called melspectrogram snippets, 2) thereafter a pretrained YAMNet architecture is employed which helps to classify these snippets into five classes for multi-class COPD severity categorization. Based on the experimental analysis using lung sounds from RespiratoryDatabase@TR [34], the proposed framework achieves a high classification performance for both two-class and multi-class COPD severity classification, which can be intuitively understood from Fig. 1(b). It can be clearly observed that all the lung sound data from five different classes are well separated in the feature plane, after using the proposed framework. This proves the superiority of the proposed framework. To the best of our knowledge, this is the second work after Altan et al. [18], [19], focusing on improving the accuracy in both two-class and multi-class COPD severity classification. The salient contributions of this paper are as follows:

- We have introduced a novel melspectrogram snippet representation learning framework, to classify different severity levels of COPD using lung sound signals. To the best of our knowledge, this is the second work on lung sound based COPD severity classification.
- For the first time, we have investigated the potential of melspectrogram time frequency representation (TFR) of lung sound signal for COPD severity classification.
- We have exploited the potential of transfer learning by fine-tuning the state-of-the-art (SOTA) pretrained audio classification network: YAMNet [32] which is explicitly pretrained using audio TFRs, for efficient COPD severity classification.
- Utilizing the SOTA network, we have achieved high classification performance of 99.25% and 96.14% for two-class and multi-class severity classification respectively. Therefore, our proposed framework outperforms existing methods in [18], [19], which are the only existing works proposed by Altan et al.
- Evaluation of the proposed framework on the RespiratoryDatabase@TR dataset by splitting the dataset into 80%-20% non-overlapping training and testing subsets respectively for overall and channel-wise performance unlike only overall performance in [18], [19].

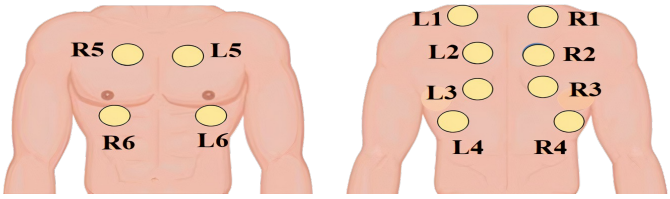

Fig. 2. Lung auscultation positions on the anterior and posterior side of body (L1-L6, R1-R2).

The rest of this paper is organized as follows: Section II discusses the public dataset used to evaluate the proposed framework. The main processing steps of the proposed framework is presented in Section III. Section IV interprets the performance evaluation, compares the test results with some of the notable prior studies and Section V concludes the proposed framework.

II. Description of Dataset

RespiratoryDatabase@TR [34] is a public multimedia respiratory dataset. For each patient, the dataset contains 4-channel phonocardiogram signals, 12-channel respiratory sounds, and spirometric measurements. The right (R) and left (L) channel lung sound signals were collected by using a Littmann 3200 digital stethoscope with the assistance of two experienced pulmonologists, from six different places of the lung auscultation region. Fig. 2 illustrates the lung auscultation positions from both the posterior and anterior sides of the body. The individuals were labeled based on the wheezing features of respiratory sounds and spirometric measurements. The current stage of COPD severity for each subject was approved by two pulmonologists, who also agreed on the diagnosis. The dataset comprises lung sounds from 41 COPD patients with varied degrees of severity, starting from COPD0 to COPD4. Five of the 41 individuals had COPD0, five had COPD1, seven had COPD2, seven had COPD3, and seventeen had COPD4. The demographics, gender, age, and information related to the auscultation process are all covered in [34]. The subjects were asked to cough at the start of each lung sound recording to ensure that the signals from the right and left channel of both the lung area were synchronized. Each recording lasted for at least 17 seconds. Lung sound signals were sampled at 4000Hz. While recording the data using the digital stethoscope, the authors of the dataset were able to remove 85% of the background noise from the room at which auscultation was done.

III. Proposed Framework

As mentioned earlier, the proposed framework consists of following steps: preprocessing, melspectrogram snippet generation and classification using pre-trained deep learning models.

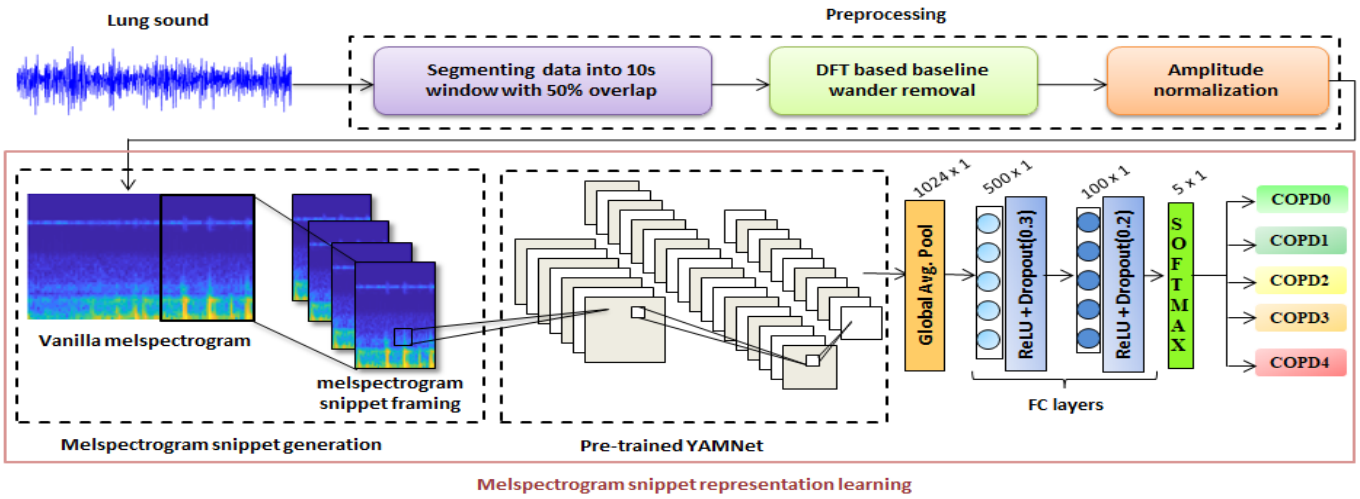

Fig. 3. Block diagram of the proposed framework for five class severity classification.

The block diagram of the proposed framework has been shown in Fig. 3. The blocks used in the proposed framework are explained in the following subsections.

A. Pre-processing

To avoid the loss of key lung sound components, all the lung sound audio signals are re-sampled at 16 kHz [13]. Basic preprocessing of lung sound signal, includes segmentation, baseline wander removal and amplitude normalisation. The respiratory sounds in the RespiratoryDatabase@TR [34] dataset are not consistent in length, with lung sound signals lasting at least 17 seconds. To draw certain pathological information and status about respiratory health, it is recommended to auscultate for one or more respiratory cycles (inspiration to expiration phase) in each auscultation site [35]. The average time for completing a respiratory cycle is approximately 5 sec [35]. In this work, we have considered 10 sec window length which covers roughly 2 cycles, and can capture significant of information about the lung sound signal which may be beneficial for classification task. Therefore, the lung sounds are splitted into 10 sec segments with 50% overlap to keep the uniform processing length. Following that, we have removed the baseline wandering (BW) component from the segmented signal, which employs a discrete Fourier transform (DFT) based filtering operation [36]. Let $s(m)$ be a lung sound signal with $m = 0, 1, 2, 3, ..., M-1$, where, M indicates the total number of samples. The DFT of the lung sound signal $s(m)$ is computed as:

$$S(p) = \sum_{m=0}^{M-1} s(m) e^{\frac{-j2\pi mp}{M}} \tag{1}$$

Generally, frequency range of BW component lies between 0 to 1 Hz. Thereby, we can eliminate the BW component from the DFT coefficients by removing the frequencies which are smaller than 1 Hz. For the f Hz frequency component, the DFT coefficient index p can be computed as follows: $p = \lceil \frac{fM}{f_s} \rceil$ where, f_s denotes the sampling rate of the lung sound signal. The BW component eliminated signal, $v(m)$ can be computed as [36]

$$v(m) = \frac{1}{M} \sum_{p=0}^{M-1} \tilde{S}(p) e^{\frac{j2\pi mp}{M}} \tag{2}$$

The thresholded DFT coefficients, denoted by $\tilde{\mathbf{S}}$, is produced as $\tilde{S}(p) = [0, ..., 0, S[p+1], ...S[M-p-1], 0, ..., 0]$. Thereafter, the BW component removed signal $v(m)$ is normalized.

$$s_n(m) = \frac{v(m)}{max|(v(m))|} \tag{3}$$

where, $s_n(m)$ indicates the normalized signal. Fig. 4 illustrates the original lung sound signal: $s(m)$ in panel (a), extracted baseline wander component: $s(m) - v(m)$ in panel (b), baseline wander removed signal: $v(m)$ in (c) and finally the normalised lung sound signal: $s_n(m)$ in panel (d).

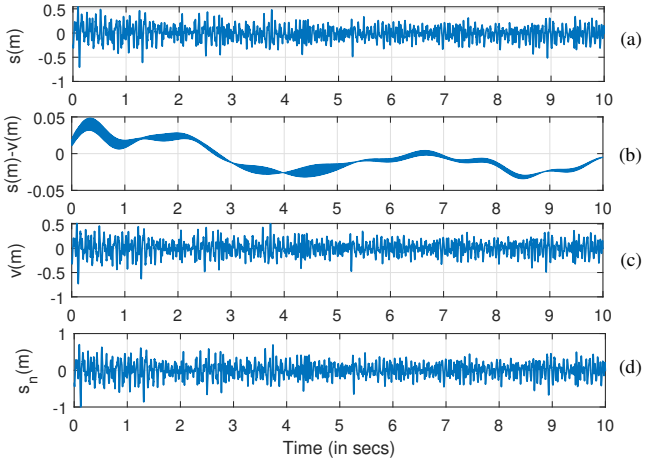

Fig. 4. (a) Original lung sound signal, (b) Baseline wander component, (c) Baseline wander removed signal, (d) Normalised lung sound signal.

B. Data augmentation

In the RespiratoryDatabase@TR COPD4 resembles as the majority class and rest classes belong to the minority class as each of these classes contain lesser subjects. To deal with

this class imbalance problem, we have adopted different time domain audio data augmentation techniques such as:

1) Time stretching: The audio sample can be speed up or slow down while maintaining pitch [37]. In this study, each of the minority class signals were stretched by two factors: $\{0.4,0.17\}$.

2) Pitch shifting: Pitch shifting modifies the pitch of the audio signal either by raising or lowering the pitch, while the duration of the audio signal is kept unaltered. In [38], the importance of pitch shifting process is investigated for CNN - based environmental sound classification. we have used two semitones or pitch shifting factors of $\{-2,1\}$ in order to augment the minority class recordings for this study.

3) Noise addition: We have also used white noise addition as another audio data augmentation strategy to increase the number of sample in the minority class.

C. Melspectrogram snippet representation learning

STFT [39], [15], scalogram [40], melspectrogram [16], [13], gammatone spectrogram [13], [14], constant Q transform (CQT) spectrograms [13], [14] have been one of the most popular input representations for lung sound classification problem. In this paper, we have proposed a novel melspectrogram snippet representation learning framework for COPD severity classification as several researches show that melspectrograms performs better than other time frequency representations (TFR) in classifying audio signals [41].The snippet representation learning consists of two sub modules: **firstly**, the lung sound are converted to vanilla melspectrogram and thereafter framed into sub-images, called melspectrogram snippets, and **secondly**, thereafter, these frames or the mel-snippets are fed as individual instances to the deep learning model YAMNet which generates feature vector corresponding to each snippet.

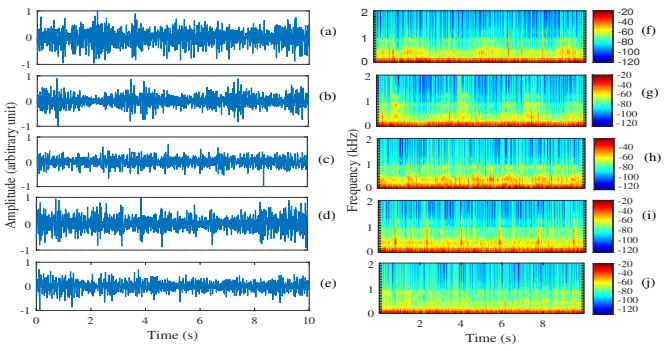

Fig. 5. (a-e) Time domain visualization of COPD0-4 signal; (f-j) Vanilla melspectrogram representation of corresponding time domain signal.

1) Melspectrogram snippet generation: In this study, initially the preprocessed signal is transformed into spectrogram with 256 frequency bins. To convert the signal into spectrogram, we have used STFT with 25 ms window, an overlap of 10 ms, and periodic Hanning window. These above mentioned values are taken similar to the original implementation of YAMNet [32], for large-scale audio classification proposed by Hershey et al. [42]. These spectrograms are then transformed to a melspectrograms with mel bins of 64. The log-melspectrograms are computed by taking the natural log of the offset melspectrogram, where the offset melspectrograms are created by adding an offset of 0.01, to avoid computing the logarithm of 0 [43]. The log-scaled melspectrograms are referred as vanilla melspectrogram, which are framed into sub-images of 0.96 s, where the frame hop is considered as 0.096 s. These framed sub-images are referred to as melspectrogram snippets, having size of 96×64 each and fed into the pretrained YAMNet, which generates feature vectors for each snippet. Fig. 5(a-e) illustrates time domain visualization of COPD0, 1, 2, 3, 4 signals respectively and Fig. 5(f-j) depicts the vanilla melspectrogram representation of corresponding time domain signals. It can be observed from the figures that vanilla spectrograms demonstrate the spectral component variation for different severity levels of COPD.

2) Snippet representation learning with pre-trained YAMNet: Training a deep convolution neural network (DCNN) from scratch requires huge volume of data. Since the acquiring lung sound data of different COPD severity levels is quite challenging due to unavailability of frequent COPD patients and expert annotations, transferring knowledge from pretrained networks, that have been trained on extensively large audio dataset, is quite useful. As a result, when such pretrained DCNN is fine tuned to accomplish some other target task, less data is required, which allows faster learning and improved performance after model fine tuning. In recent years, several works has been carried out on lung sound anomaly classification, using standard pretrained DCNN models like ResNet50 [39], ResNet34 [15] via transfering their knowledge. These DCNN models are initially trained using images from ImageNet database [44]. In [39], [15] the lung sound signal is converted to spectrogram images and fed to these DCNN models which acts as deep feature extractor. However, the final classification results are degrading as these DCNN are pretrained with images. It has been investigated that as spectrograms contain immense information of audio signal it is difficult to extract promosing feature from image based CNN models [45]. For this context Tsalera et al. [45] has investigated that, if we want to use knowledge transfer technique on sound classification problem, it will be fruitful to transfer knowledge from such DCNN which has already been pretrained with audio TFRs such as melspectrogram and spectrograms rather that simple images. Thereby, in recent years research on audio neural networks has received significant attention [32], [46]. Hence, in this work, we have used a pretrained audio classification model: YAMNet [32], [46]. YAMNet has been trained on the melspectrograms extracted from the audio signals of AudioSet [33] which is the largest dataset for audio deep learning. Let us consider, a deep learning model M_0 was pretrained using the source dataset $P_s = \{a_s^i, o_s^i\}_{i=1}^{n_s}$, where, P_s contains input features (a_s^i) and corresponding labels (o_s^i). On the target dataset $P_t = \{a_t^i, o_t^i\}_{i=1}^{n_t}$, we aim to fine tune the pretrained model using the transfer learning approach in order to generate better results on the target task. In this work, the source dataset is AudioSet and the target dataset is RespiratoryDatabase@TR [34]. At

the time of fine-tuning, only P_t and M_0 are available. As P_s and P_t are of different domains and may have different input and output spaces, we can not use M_0 directly to the target data. Generally, DCNNs are often divided into two portions: a generic representation function $F_{\bar{\lambda^0}}$ (parameterized by $\bar{\lambda^0}$) and a domain or task specific function G_{λ_s}, which is represented by the the top layers of the DCNN. In case of transfer learning based approach, the generic representation function is retained, while the domain or task specific function is replaced by randomly initialised functionality D_{λ_t} (paramerterized by λ_t). Therefore, we optimize:

$$\left(\bar{\lambda}^*, \lambda_t^*\right) = \underset{\bar{\lambda}, \lambda_t}{\operatorname{argmin}} \sum_{i=1}^{n_t} \mathscr{L}\{D_{\lambda_t}\{F_{\bar{\lambda}}\left(a_t^i\right)\}, o_t^i\} \tag{4}$$

where, $\mathscr{L}\{\cdot\}$ indicates the loss function. At the commencement of the optimization process, these pretrained parameters $\bar{\lambda^0}$ provide good starting point. YAMNet is built on the MobileNet_v1 architecture and is made up of depth-wise separable and point-wise convolutions, which significantly helps in reducing model size and computing cost. The architecture of YAMNet comprises of one convolutional layer, 13 pairs of depth-wise separable and point-wise convolution layers. After each convolution layer, ReLU activation and batch normalization is used. Lastly, a global average pooling (GAP) layer is applied along with a fully connected classifier layer to classify the audio signals. The top layers of the YAMNet have been removed for fine-tuning purposes, resulting in a feature vector of size 1024 as the new output of the YAMNet, which can be achieved from the GAP layer of YAMNet. Thereafter, the feature vector is passed through two fully connected (FC) layers or dense layers, consisting 500 and 100 neurons respectively. The activation function used for these FC layers is ReLU. Dropout rate of 0.3 and 0.2 are used in the FC layers respectively, to alleviate overfitting problem. Then, dense layer output is fed to the classification layer that generates $C \times 1$ dimensional class output, where C indicates the total number of classes (where, $C = 2$ for binary and $C = 5$ for multi-class COPD severity classification). In general, the operation for the classification layer can be represented as:

$$Output = \sigma(< X_{dense}, W_0 > + b_0) \tag{5}$$

where, $< X_{dense}, W_0 >$ denotes the dot product between weight vector (W_0) and the output of the dense layer (X_{dense}), b_0 denotes the bias and σ refers to the activation function. For binary and multi-class severity classification, we have employed the sigmoid and softmax activation functions, respectively. The sigmoid function is given by [47]:

$$\text{Sigmoid}(y) = \frac{1}{1 + e^{-y}} \tag{6}$$

Softmax activation function can be expressed as [48]:

$$\text{Softmax}(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{C} e^{y_j}} \tag{7}$$

where y_i is the i^{th} element from the output of neural network and the numerator is normalised to bring the class probability value between 0 to 1.

IV. Results and Discussion

In this section, the performance of the proposed framework is examined by utilizing publicly available multimedia RespiratoryDatabase@TR using different widely used performance metrics which are presented in the subsequent subsections:

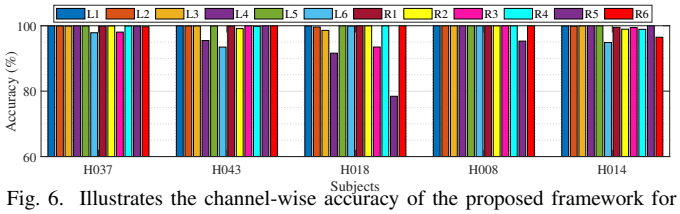

Fig. 6. Illustrates the channel-wise accuracy of the proposed framework for multi-class COPD severity classification.

A. Performance metrics

To evaluate our proposed framework, we have used the following performance metrics: accuracy (ACC), precision (PRC), sensitivity (SEN)/recall (REC), specificity (SPE), and Mathew's correlation coefficient (MCC) for both binary and multiclass classification similar to [49], [50]. These metrics can be computed from the confusion matrix [51]. Further for multi-class severity classification we have used two new performance metrics, namely weighted specificity (WSPE), and weighted sensitivity (WSEN) [19]

B. Performance evaluation

The performance of the proposed framework is assessed via training and testing mechanism. Using the proposed framework, we have accomplished both binary and multi-class severity classification. The optimized simulation parameters are obtained after extensive training of the proposed DCNN architecture, which are provided in Table I. We have evaluated our framework through two experiments. In **experiment 1:**, we have splitted the entire dataset (patient-wise) into 80%-20% non-overlapping train and test subsets, respectively, which helps to prevent the scenario in which lung sound signals from one individual remains in both the train and test subset [13]. For this experiment, we have presented the channel-wise accuracy for different subjects present in the test subset. Fig. 6 and Table II illustrates the channel-wise or acquisition site-wise average accuracy for five class severity classification after performing experiment 1, which proves that our proposed framework can accurately identify different severity levels from all the 12-channel lung sound signals with high classification rate. Whereas in **experiment 2**, we have randomly splitted the entire dataset into 80%-20% ratio for training-testing. Further, 20% data for testing is splitted into 10% each for testing and validation, similarl to [49], [50]. For this experiment, we have presented overall training and testing performances through loss, accuracy curves, and other performance metrics. A batch size of 128 was utilized to train the deep learning model, coupled with the Adam optimizer. A fixed valued learning rate of 0.0003 was used to fine-tune and train the DCNN model. We have applied five fold cross-validation to ensure the generalization of our proposed framework. We have utilized binary and categorical cross-entropy loss for two class and multi-class severity classification tasks, respectively. Fig. 7(a-b) depicts the training and

validation accuracy, loss graphs of the proposed deep learning framework for binary severity classification, and Fig. 7(c-d) shows the accuracy and loss graphs for multi-class severity classification. From the Fig. 7 it can be observed that the deep learning framework has learned the promising class-dependent features from the melspectrogram representations and does not overfit. As we have used fine tuning on the pretrained network, we can see that within less number of epochs the network has achieved high classification accuracy for both binary and multi-class severity classification. Table III and IV illustrates the fold-wise classification performance obtained using the proposed framework for both binary and multiclass severity classification. Different classification metrics have been tabulated for each of the five folds. It can be observed from both Table III and IV that our proposed framework provides outstanding results for each of the fold which in a way, refers to the generalization of the proposed framework and outperforms the existing results on both binary [18] and multiclass severity classification task [19].

TABLE I
SIMULATION PARAMETERS

Parameter	Details
Size of input snippet	96×64
Fully connected layers	2 (Neurons: 500, 100)
Dropout factor	0.3 ,0.2
Optimizer	Adam
Learning rate	0.0003
Batch size	128

TABLE II
CLASSIFICATION ACCURACY WITH RESPECT TO 12 DIFFERENT ACQUISITION SITES (L1-L6, R1-R6)

	Classification accuracy of subject no. (COPD severity type) with respect to acquisition sites						
	H037 (COPD0)	**H043 (COPD1)**	**H018 (COPD2)**	**H008 (COPD3)**	**H0026 (COPD3)**	**H014 (COPD4)**	**H006 (COPD4)**
L1	0.993	0.99	0.993	0.991	0.865	0.993	0.894
L2	0.99	0.99	0.998	0.99	0.899	0.998	0.889
L3	0.99	0.995	0.975	0.984	0.899	0.993	0.887
L4	0.997	0.952	0.914	0.987	0.842	0.99	0.907
L5	0.99	0.994	0.984	0.993	0.869	0.99	0.832
L6	0.978	0.934	0.987	0.99	0.87	0.948	0.785
R1	0.998	0.997	0.993	0.99	0.894	0.995	0.702
R2	0.99	0.981	0.989	0.948	0.905	0.981	0.933
R3	0.971	0.99	0.924	0.989	0.891	0.94	0.793
R4	0.998	0.99	0.995	0.991	0.782	0.989	0.814
R5	0.993	0.991	0.785	0.953	0.945	0.991	0.803
R6	0.99	0.99	0.998	0.989	0.944	0.968	0.884

TABLE III
FOLD-WISE CLASSIFICATION RESULTS OBTAINED FOR BINARY SEVERITY CLASSIFICATION

Performance metrics (%)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
ACC	99.35	99.18	99.25	99.25	99.26	99.25
SEN	99.25	99.1	99.18	99.19	99.18	99.18
SPE	99.4	99.36	99.36	99.32	99.35	99.36
PRC	99.25	99	99.2	99	99	99.09
F1-Score	99.15	99.13	99.11	99.13	99.13	99.13

From Fig. 8, it can also be argued that our model is capable of classifying the different severity levels accurately as the features corresponding to each class make distinct clusters in the 2D-feature plane, which has been visualized by t-SNE visualization plots for both binary and multi-class classification. Table V illustrates the class-wise performance analysis of the multiclass severity classification task. From the confusion matrix obtained using the test data while evaluating the framework, we can calculate all possible performing metrics for each of the classes [51]. Table V contains classwise SEN, SPE, ACC, PRC, and F1-score values for each of the severity classes present in the multiclass classification. It can be observed from the table that the proposed framework classifies each of the severity classes with almost 97% accuracy, and other metrics such as F1-score of each of the class also yields high value. This refers to the supremacy of the proposed framework.

TABLE IV
FOLD-WISE CLASSIFICATION RESULTS OBTAINED FOR MULTICLASS SEVERITY CLASSIFICATION

Performance metrics (%)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
ACC	96.16	96.15	96.14	96.14	96.15	96.14
W-SPE	96	95.93	95.92	95.92	95.93	95.94
W-SEN	98.99	98.89	98.88	98.88	98.89	98.89

TABLE V
CLASSIFICATION PERFORMANCES (%) OF EACH COPD SEVERITY FOR MULTICLASS CLASSIFICATION USING PROPOSED FRAMEWORK

COPD severity	ACC	REC	PRC	F1-score
COPD0	97.66	97.49	98.85	98.16
COPD1	97.82	98.74	96.98	97.85
COPD2	95.38	97.1	97.02	97.05
COPD3	97.22	97.06	97.66	97.35
COPD4	97.93	96.98	98.95	97.96

TABLE VI
COMPARATIVE ANALYSIS OF PROPOSED FRAMEWORK WITH DIFFERENT TFRS FOR BINARY AND MULTICLASS SEVERITY CLASSIFICATION

Metric (%)	Type of time frequency representation			
	STFT Spectrogram	**Mel spectrogram**	**Gammatone spectrogram**	**CQT spectrogram**
ACC (Binary)	96.99	99.25	95.62	92.47
ACC (Multiclass)	93.01	96.14	93.6	90.53

C. Analysis sensitiveness of the proposed framework

1) Effect of processing length of input lung sound signal: Since lung sound signals are highly non-stationary signals, we generally process TFR of the lung sound signal. However, the information of the lung sound time series is present in a sequential manner, the temporal resolution or processing length of the input signal also influences the classification performance of deep learning model. We investigated three different processing lengths ranging from 5 sec to 15 sec and carried out the binary, multiclass COPD severity classification task to assess the influence of processing length. Fig. 9(a-b) illustrates the variation of classification parameters of the proposed framework with respect to different input length of lung sound signal for both the tasks. With different experimentation, we have found that lung sound with 10 sec processing length yields highest classification rates for both the classification task. While 15 sec processing length is also yielding competitive results with respect to 10 sec, however slight reduction in the metrics can be observed. Additionally we can observe that with shorter segment length i.e., 5 sec processing length the performance degrades drastically. According to Sarkar et al. [35], we should use at least two respiratory cycle for assess any lung sound signal. As 5 sec. frame just covers only one

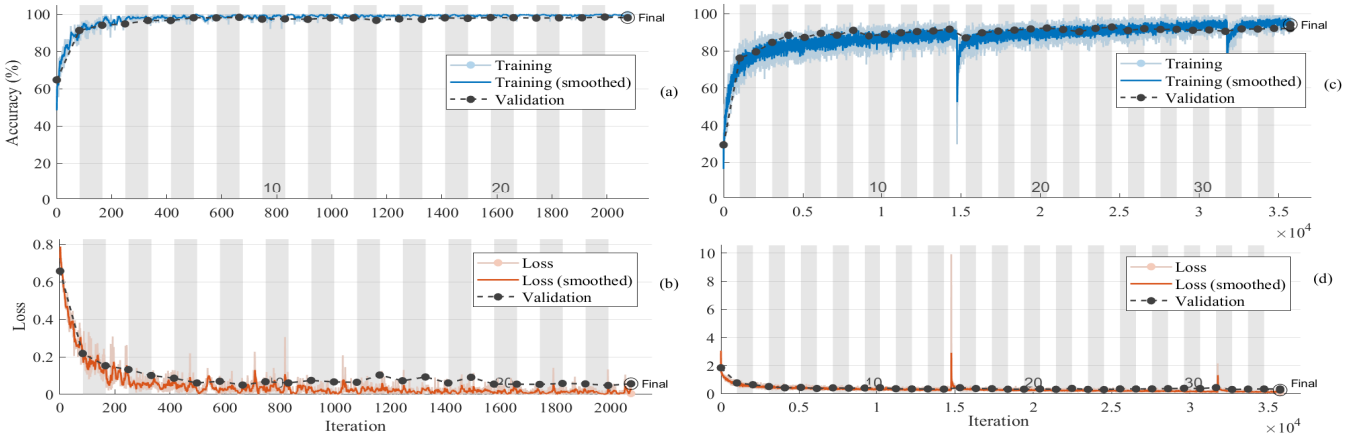

Fig. 7. (a-b) Classification accuracy and loss plots for binary COPD severity classification, (c-d) Classification accuracy and loss plots for multi-class severity classification.

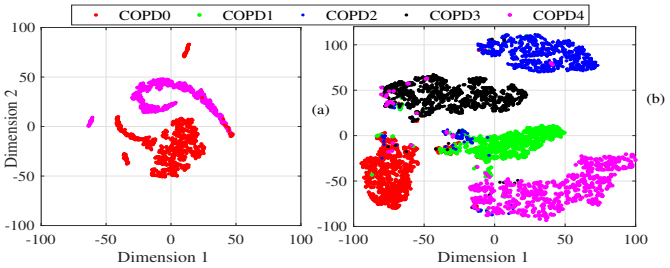

Fig. 8. Two-dimensional visualization of high-dimensional features at the first fully connected layer of the testing observations for (a) Binary severity classification and (b) Multiclass severity classification.

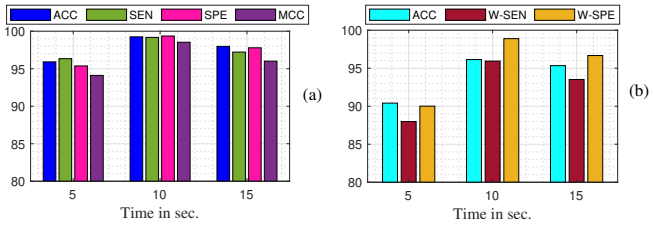

Fig. 9. Effect of processing length of input lung sound signal on different classification metrics for (a) binary severity classification and (b) multiclass severity classification

respiratory cycle (approx.) therefore, it has substantially lesser information; that's why we can also see the degradation in classification performance in Fig. 9(a-b).

2) Choice of proper time - frequency representation: Lung sound signals are highly non-stationary signals. Thereby, it is necessary to transform the signal from one domain to another in order to gain a comprehensive understanding of them. In this regard, TFRs captures several spectral variations of the lung sound signal over time. Hence, choice of TFR also affects classification accuracy of the the DCNN model. Therefore, in Table VI, a comparative result analysis is presented based on the effect on classification accuracy with respect to different TFRs. In this research, we have mainly stressed upon the mel-spectrogram representation which has already been established as one the best TFR for audio signal classification task [41] as it captures wealth of time-frequency information from the audio signal [41]. From Table VI, it can be observed that with the use of the snippets of melspectrogram the classification accuracy reaches the maximum value for both binary and multiclass COPD severity classification. With the use of simple STFT based spectrogram and gammatone spectrogram snippets the classification accuracy degrades quite a bit. However, CQT spectrogram snippets performs the worst among all other TFRs in classifying the COPD severities.

TABLE VII
COMPARATIVE ANALYSIS OF YAMNET BASED FINE-TUNING FRAMEWORK WITH OTHER DEEP LEARNING MODELS

Model evaluation parameter	VGG 16	Alex Net	ResNet 50	Mobile NetV2	YAM Net
Total trainable parameters (M)	138	25.7	25.6	4.2	3.8
Model size	1.5 GB	294 MB	98MB	49 MB	13 MB
ACC (%) (2-class)	92.62	91.65	93.26	87.82	99.25
ACC (%) (5-class)	84.38	82.95	76.03	89.36	96.14

TABLE VIII
PERFORMANCE (%) COMPARISON FOR BINARY CLASSIFICATION

Author	ACC	SEN	SPE	PRC	MCC	F1-score
Altan et al. [18]	95.85	93.34	93.65	—	—	—
Proposed Framework	99.25	99.18	99.36	99.09	98.53	99.13

TABLE IX
PERFORMANCE (%) COMPARISON FOR MULTICLASS CLASSIFICATION

Author	ACC	W-SEN	W-SPE
Altan et al. [19]	94.31	94.28	98.76
Proposed Framework	96.14	95.94	98.89

3) Choice of proper deep learning model for transferring knowledge for COPD severity classification: In this part, we have experimented with different well known DCNN models for COPD severity classification using knowledge transfer technique. We have compared our proposed framework with VGG-16 [52], AlexNet [53], Resnet50 [29] and MobileNet-V2 [54] in terms of model size, total trainable parameters, obtained accuracy for both binary and multiclass COPD severity classification. The comparison results are tabulated in Table VII. From Table VII, it can be observed that our proposed approach outperforms the other ImageNet [44] pre-trained DCNNs [52], [53], [29], [54]. Proposed framework employs the potential of transferring knowledge from the audio neural network: YAMNet. To prove the better domain

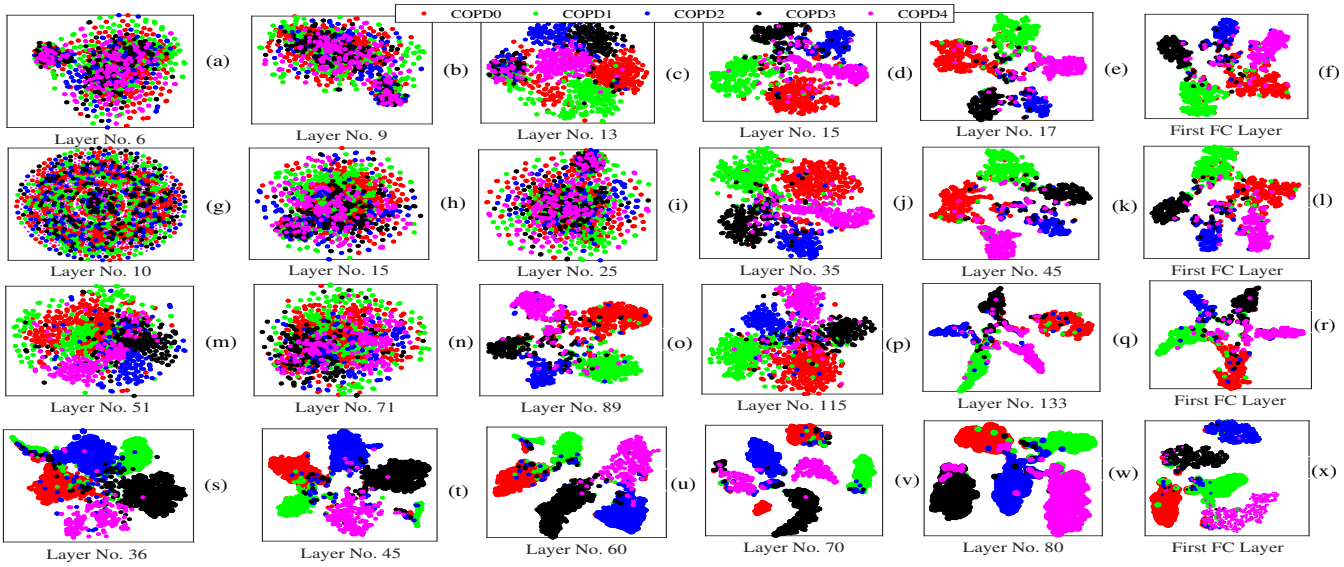

Fig. 10. t-SNE plots of extracted features from different layers of (a-e) finetuned VGG-16, (g-l) finetuned ResNet-50, (m-r) finetuned MobileNetV2, (s-x) finetuned YAMNet.

TABLE X
PERFORMANCE COMPARISON WITH EXISTING WORKS ON COPD SEVERITY DETECTION

Reference	Data Type	Features	Classifiers	Classes	ACC(%)	SEN(%)	SPE(%)	W-SEN(%)	W-SPE(%)
Moghadas-Dastjerdi et al. [55]	CT images	Image Features	Naive Bayes	4	84.18	82.64	—	—	—
Ying et al. [5]	Clinical data	Spirometry,questionnaires	DBN	4	97.20	—	—	—	—
Newandee et al. [20]	Heart rate variability	PCA based feature	Clustering Analysis	5	88.00	—	—	—	—
Morillo et al. [25]	Trechial sound	STFT	ANN	2	77.60	72.00	81.80	—	—
Sugimori et al. [28]	CT images	Raw CT images	ResNet50 CNN	5	44	—	—	51	—
Du et al. [27]	CT image	Airway tree feature	Deep mult-iview CNN	2	88.6	96.3	—	—	—
Ho et al. [26]	CT image and PRM	Parenchymal variables	3D-CNN model	2	89.3	88.3	—	—	—
Altan et al. [18]	Lung sound [34]	3D-SODP features	DBN	2	95.85	93.34	93.65	—	—
Altan et al. [19]	Lung sound [34]	3D-SODP features	Deep ELM	5	94.31	—	—	94.28	98.76
Proposed Framework	**Lung sound** [34]	**Melspectrogram**	**Finetuned YAMNet**	2	**99.25**	**99.18**	**99.36**	—	—
				5	**96.14**	—	—	**95.94**	**98.89**

adaptation quality of YAMNet, we have provided a group of t-SNE plots which are created using the features extracted from intermediate layers of fine tuned VGG-16, fine tuned ResNet-50, fine tuned MobileNetV2, and fine tuned YAMNet depicted in Fig. 10 (a-f), Fig. 10 (g-l), Fig. 10 (m-r), and Fig. 10 (s-x) respectively. It can be clearly observed that the intermediate layers of the fine tuned YAMNet model has better domain adaptation quality than any other pre-trained networks, as it yields good enough t-SNE feature clusters in each of the intermediate layers to that of the other networks. This establishes the fact that, using an audio pre-trained network for lung sound audio classification tasks offers higher knowledge transfer capabilities than the traditional image pre-trained models [54], [29], [52]. This also signifies that, pre-trained YAMNet weights provides good weight initialization for the lung sound based COPD severity classification task as it can be observed from the Fig. 10(s-x) that intermediate layers of YAMNet outperforms all other pre-trained networks by extracting class independent features (which helps to form better clusters in the t-SNE feature plane). In a way, the results of Table VII proves that, since YAMNet is explicitly pre-trained by the melspectrograms of audio signals, therefore, it is beneficial to use domain adaptability nature of YAMNet by transferring knowledge on lung sound based COPD severity classification task rather than using other ImageNet pre-trained DCNNs [45]. Hence, choosing YAMNet over other DCNN model is justified. Additionally, our suggested framework performs better in terms of classification accuracy for both tasks than the so-called lightweight DCNN model MobileNet-v2 by a margin of 11.43% and 6.78%, offering a well trade-off between the amount of parameters, using substantially less storage space. We have computed the total execution time required to classify one lung sound signal using our proposed framework. It takes nearly 1.221 ± 0.02 seconds to classify a lung sound signal at MATLAB 2021b 64 bit OS on Windows 10, 32GB RAM desktop consisting Intel Xeon(R) W-1350 processor with clock frequency of 3.30 - 3.31 GHz. The preprocessing part takes 0.083 seconds, mel-spectrogram snippet generation process requires approximately 0.103 ± 0.03 seconds and finally the YAMNet based final classification takes 1.035 ± 0.02 seconds to execute. However, the Mobilenet-V2 needs 1.1 ± 0.01 seconds to classify the same melspectrogram snippets. Therefore, it also proves that our proposed framework achieves SOTA performance in terms of classification accuracy, total number of trainable parameters and execution time.

D. Performance comparison

In this subsection, the superiority of our proposed framework is analyzed with respect to other existing COPD severity classification techniques. Even though clinical studies support

that lung auscultation is one of the most unique diagnostic techniques for detecting different respiratory disorders, the majority of prior research on computer-assisted analysis has emphasized the clinical, pathological data, and various other biological signals. Till date, very few research has been carried out that focuses on lung sounds for COPD severity detection. To the best of our knowledge, this is the second work that focuses on both binary and multi-class COPD severity classification based on lung sounds after Altan et al [19], [18]. Hence, first we compare our performance results with the methods proposed by Altan et al. Table VIII and Table IX depict the performance results in comparison with the method proposed by Altan et al. [19] for binary classification (COPD0 and COPD4) and multi-class COPD severity classification respectively. It is clearly observed from the tables that proposed framework provides higher classification performance in terms of all evaluation metrics.

Also, in [19], low classification accuracy rates were achieved for COPD1 and COPD3 which is significantly higher using proposed framework. We also perform a detailed comparative performance analysis for COPD severity categorization using other biomedical modalities along with the lung sounds. Table X, represents the overall comparative study by considering different datasets, methods, classifiers, and performance measures. It can be observed from the table that existing works exploited the use of subjective measurements including symptoms, medical assessments, physical examinations, questionnaire responses [5] and different biomedical input modalities including CT scans [55], [28], [27], PRM method [26], spirometric measures [5], electrocardiogram (ECG) [20] and tracheal sounds [25] for COPD severity classification. However, the proposed framework exploits the potential of time-frequency melspectrogram representation of lung sound signals and a pretrained audio classification network i.e., YAMNet, which help to achieve the highest classification performance both in binary classification and multi-class classification for COPD severity as compared with existing works.

V. Conclusion

In this paper, we have proposed a melspectrogram snippet representation learning framework for COPD severity classification. This works exploits the time-frequency melspectrogram snippets representation of lung sounds and YAMNet based transfer learning model. The proposed framework works in following stages: preprocessing, melspectrogram snippet representation generation from lung sound and fine tuning of a pretrained YAMNet. The proposed framework achieves an accuracy of 99.25% and 96.14% for binary and multi-class COPD severity classification respectively for lung sounds taken from publicly available RespiratoryDatabase@TR dataset. Further, experimental results demonstrate that the proposed framework achieves superior classification performance as compared with existing works for COPD severity classification. However, in the future, we hope to investigate the impact of various metrological influences on COPD severity classification performance, such as ambient noise and heart sound interference, and how to eradicate them in order to achieve better results. At the same time, we believe that our approach will make it possible to create a system that can automatically detect COPD severities from lung auscultations in actual clinical settings.

References

[1] A. A. Cruz, *Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach*. World Health Organization, 2007.

[2] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Medicine*, vol. 3, no. 11, p. e442, 2006.

[3] S. Guessoum, M. T. Laskri, and J. Lieber, "Respidiag: a case-based reasoning system for the diagnosis of chronic obstructive pulmonary disease," *Expert Systems with Applications*, vol. 41, no. 2, pp. 267–273, 2014.

[4] V. Cheplygina, L. Sørensen, D. M. Tax, J. H. Pedersen, M. Loog, and M. de Bruijne, "Classification of copd with multiple instance learning," in *2014 22nd International Conference on pattern recognition*. IEEE, 2014, pp. 1508–1513.

[5] J. Ying, J. Dutta, N. Guo, L. Xia, A. Sitek, Q. Li, and Q. Li, "Gold classification of copdgene cohort based on deep learning," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2474–2478.

[6] A. B. Bhome, "Copd in india: Iceberg or volcano?" *Journal of thoracic disease*, vol. 4, no. 3, p. 298, 2012.

[7] K. G. Fan, J. Mandel, P. Agnihotri, and M. Tai-Seale, "Remote patient monitoring technologies for predicting chronic obstructive pulmonary disease exacerbations: review and comparison," *JMIR mHealth and uHealth*, vol. 8, no. 5, p. e16147, 2020.

[8] N. Meslier, G. Charbonneau, and J. Racineux, "Wheezes," *European Respiratory Journal*, vol. 8, no. 11, pp. 1942–1948, 1995.

[9] A. Alvar, M. Decramer, and P. Frith, "Global initiative for chronic obstructive lung a guide for health care professionals global initiative for chronic obstructive disease. global initiative for chronic obstructive lung disease, 22 (4), 1–30," 2010.

[10] N. I. of Health *et al.*, "National asthma education and prevention program," *Expert panel report*, vol. 3, 1997.

[11] S. Jayalakshmy and G. F. Sudha, "Scalogram based prediction model for respiratory disorders using optimized convolutional neural networks," *Artificial Intelligence in Medicine*, vol. 103, p. 101809, 2020.

[12] S. Vadrevu and M. S. Manikandan, "A robust pulse onset and peak detection method for automated ppg signal analysis system," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 3, pp. 807–817, 2018.

[13] 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*, 2021.

[14] D. Ngo, L. Pham, A. Nguyen, B. Phan, K. Tran, and T. Nguyen, "Deep learning framework applied for predicting anomaly of respiratory sounds," in *2021 International Symposium on Electrical and Electronics Engineering (ISEE)*. IEEE, 2021, pp. 42–47.

[15] 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 *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 527–530 .

[16] 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)*, 2020, pp. 760–763.

[17] G. Altan, Y. Kutlu, and N. Allahverdi, "Deep learning on computerized analysis of chronic obstructive pulmonary disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 5, pp. 1344–1350, 2019.

[18] G. Altan, Y. Kutlu, A. Ö. Pekmezci, and S. Nural, "Deep learning with 3d-second order difference plot on respiratory sounds," *Biomedical Signal Processing and Control*, vol. 45, pp. 58–69, 2018.

[19] G. Altan, Y. Kutlu, and A. Gökçen, "Chronic obstructive pulmonary disease severity analysis using deep learning on multi-channel lung sounds," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 28, no. 5, pp. 2979–2996, 2020.

[20] D. Newandee, S. Reisman, A. Bartels, and R. De Meersman, "Copd severity classification using principal component and cluster analysis on hrv parameters," in *2003 IEEE 29th Annual Proceedings of Bioengineering Conference*. IEEE, 2003, pp. 134–135.

[21] R. Naves, B. H. Barbosa, and D. D. Ferreira, "Classification of lung sounds using higher-order statistics: A divide-and-conquer approach," *Computer Methods and Programs in Biomedicine*, vol. 129, pp. 12–20, 2016.

[22] M. A. Fernandez-Granero, D. Sanchez-Morillo, and A. Leon-Jimenez, "An artificial intelligence approach to early predict symptom-based exacerbations of copd," *Biotechnology & Biotechnological Equipment*, vol. 32, no. 3, pp. 778–784, 2018.

[23] R. J. Oweis, E. W. Abdulhay, A. Khayal, A. Awad *et al.*, "An alternative respiratory sounds classification system utilizing artificial neural networks," *Biomed J*, vol. 38, no. 153, p. e61, 2015.

[24] J. L. Amaral, A. C. Faria, A. J. Lopes, J. M. Jansen, and P. L. Melo, "Automatic identification of chronic obstructive pulmonary disease based on forced oscillation measurements and artificial neural networks," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 1394–1397.

[25] D. Sánchez Morillo, A. Leon Jimenez, and S. A. Moreno, "Computer-aided diagnosis of pneumonia in patients with chronic obstructive pulmonary disease," *Journal of the American Medical Informatics Association*, vol. 20, no. e1, pp. e111–e117, 2013 .

[26] T. T. Ho, T. Kim, W. J. Kim, C. H. Lee, K. J. Chae, S. H. Bak, S. O. Kwon, G. Y. Jin, E.-K. Park, and S. Choi, "A 3d-cnn model with ct-based parametric response mapping for classifying copd subjects."

[27] R. Du, S. Qi, J. Feng, S. Xia, Y. Kang, W. Qian, and Y.-D. Yao, "Identification of copd from multi-view snapshots of 3d lung airway tree via deep cnn," *IEEE Access*, vol. 8, pp. 38 907–38 919, 2020.

[28] H. Sugimori, K. Shimizu, H. Makita, M. Suzuki, and S. Konno, "A comparative evaluation of computed tomography images for the classification of spirometric severity of the chronic obstructive pulmonary disease with deep learning," *Diagnostics*, vol. 11, no. 6, p. 929, 2021.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] D. M. Vasilescu, F. J. Martinez, N. Marchetti, C. J. Galbán, C. Hatt, C. A. Meldrum, C. Dass, N. Tanabe, R. M. Reddy, A. Lagstein *et al.*, "Noninvasive imaging biomarker identifies small airway damage in severe chronic obstructive pulmonary disease," *American journal of respiratory and critical care medicine*, vol. 200, no. 5, pp. 575–581, 2019 .

[31] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[32] M. Plakal and D. Ellis, "Yamnet," https://github.com/tensorflow/models/tree/master/research/audioset/yamnet, Jan. 2020.

[33] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[34] G. ALTAN, Y. Kutlu, Y. Garbi, A. Ö. Pekmezci, and S. Nural, "Multimedia respiratory database (respiratorydatabase@ tr): Auscultation sounds and chest x-rays," *Natural and Engineering Sciences*, vol. 2, no. 3, pp. 59–72, 2017.

[35] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, "Auscultation of the respiratory system," *Annals of Thoracic Medicine*, vol. 10, no. 3, p. 158, 2015.

[36] U. Satija, B. Ramkumar, and M. S. Manikandan, "Real-time signal quality-aware ecg telemetry system for iot-based health care monitoring," *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 815–823, 2017.

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

[38] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[39] R. Shethwala, S. Pathar, T. Patel, and P. Barot, "Transfer learning aided classification of lung sounds-wheezes and crackles," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1260–1266 .

[40] S. B. Shuvo, S. N. Ali, S. I. Swapnil, T. Hasan, and M. I. H. Bhuiyan, "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 .

[41] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *arXiv preprint arXiv:1706.07156*, 2017 .

[42] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135 .

[43] P. Primus, "Gradient-based explanations for audio classifiers/submitted by paul primus," Ph.D. dissertation, Universität Linz, 2019.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[45] E. Tsalera, A. Papadakis, and M. Samarakou, "Comparison of pre-trained cnns for audio classification using transfer learning," *Journal of Sensor and Actuator Networks*, vol. 10, no. 4, p. 72, 2021.

[46] 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.

[47] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018.

[48] M. Pérez-Enciso and L. M. Zingaretti, "A guide on deep learning for complex trait genomic prediction," *Genes*, vol. 10, no. 7, p. 553, 2019.

[49] M. Saini, U. Satija, and M. D. Upadhayay, "One-dimensional convolutional neural network architecture for classification of mental tasks from electroencephalogram," *Biomedical Signal Processing and Control*, vol. 74, p. 103494, 2022.

[50] E. Prabhakararao and S. Dandapat, "Multi-scale convolutional neural network ensemble for multi-class arrhythmia classification," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[51] T. Kautz, B. M. Eskofier, and C. F. Pasluosta, "Generic performance measure for multiclass-classifiers," *Pattern Recognition*, vol. 68, pp. 111–125, 2017.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[54] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520 .

[55] H. Moghadas-Dastjerdi, M. Ahmadzadeh, E. Karami, M. Karami, and A. Samani, "Lung ct image based automatic technique for copd gold stage assessment," *Expert Systems with Applications*, vol. 85, pp. 194–203, 2017.