RDLINet: A Novel Lightweight Inception Network for Respiratory Disease Classification Using Lung Sounds

Arka Roy, *Graduate Student Member, IEEE*, and Udit Satija, *Senior Member, IEEE*

***Abstract*— Respiratory diseases are the world's third leading cause of mortality. Early detection is critical in dealing with respiratory diseases, as it improves the effectiveness of intervention, including treatment and reducing the spread. The main aim of this article is to propose a novel lightweight inception network to classify a wide spectrum of respiratory diseases using lung sound signals. The proposed framework consists of three stages: 1) preprocessing; 2) mel spectrogram extraction and conversion into a three-channel image; and 3) classification of the mel spectrogram images into different pathological classes using the proposed lightweight inception network, namely, respiratory disease lightweight inception network (RDLINet). Utilizing the proposed architecture, we have achieved a high classification accuracy of 96.6%, 99.6%, and 94.0% for seven-class classification, six-class classification, and healthy versus asthma classification. To the best of our knowledge, this is the first work on seven-class respiratory disease classification using lung sounds. Whereas, our proposed network outperforms all the existing published works for six-class and binary classifications. The suggested framework makes use of deep-learning methods and offers a standardized evaluation with strong categorization capabilities. In order to distinguish between a wide range of respiratory diseases, our study is a pioneering one that focuses exclusively on lung sounds. The proposed framework can be translated into real-time clinical application, which will facilitate the prospect of automated respiratory health screening using lung sounds.**

***Index Terms*— Lightweight inception network, lung auscultation, lung sounds, mel spectrogram, respiratory disease classification.**

I. Introduction

LUNG auscultation is one of the most popular diagnostic modalities [1], [2] used by the pulmonary experts to analyze the condition of the respiratory system. When auscultating various areas on the anterior and posterior sides of the chest [3], [4], lung sounds can be detected. Lung sounds are indicative of different anatomical flaws in the lungs and provide accurate prognoses regarding respiratory health, resulting in more trustworthy medical tool for identifying respiratory disorders [5]. According to a recent study conducted by the world health organization (WHO), approximately ten million (M) people die each year as a result of respiratory diseases [6]. Another WHO study suggests that the majority of people suffering from respiratory disorders around the world may have one of the following five diseases: asthma, chronic obstructive pulmonary disease (COPD), lung cancer, tuberculosis, and lower respiratory tract infection (LRTI) [6]. Early detection is critical in dealing with respiratory diseases, as it improves the effectiveness of intervention, including treatment, and helps in restricting the spread. Even though doctors use photoplethysmograph [7], [8], spirometry [3], chest computerized tomography (CT) scan images [9], and clinical history to analyze the respiratory health, however, lung auscultation remains essential to doctors as during lung auscultation, experts can detect various adventitious respiratory sounds, such as wheeze, crackle, stridor, and so on, which indicate the presence of respiratory disorders in that individual [10]. Thereby, developing artificial intelligence (AI)-based automated algorithms will be extremely beneficial in the early detection of several respiratory diseases.

A. Related Works on Lung Sound-Based Multiclass Respiratory Disease Classification

The invention of the digital stethoscope allows for the continuous recording of lung sounds from subjects in order to detect various respiratory diseases automatically [5]. Therefore, research into the automated analysis of lung sounds has gained significant attention in recent years. The automated categorization of respiratory sounds using machine learning (ML) and deep learning has been the subject of a significant amount of prior studies. However, instead of directly predicting respiratory disorders from lung auscultation recordings, most of the approaches have concentrated on respiratory anomaly prediction, i.e., identifying the lung sounds as wheeze, crackles, and so on [11], [12], [13]. However, few research methods have been examined and assessed in recent years for automatically detecting respiratory disorders from lung auscultation sounds, which include different feature extraction processes, such as mel frequency cepstral coefficient (MFCC) [14], spectrograms [13], mel spectrograms [1], scalograms [2], and so on, followed by a wide variety of deep-learning and ML techniques. All these works have been carried out in two different resolutions: either three-class classification (chronic, nonchronic disease, and healthy) or

Manuscript received 2 March 2023; revised 17 May 2023; accepted 23 June 2023. Date of publication 6 July 2023; date of current version 17 July 2023. This work was supported by the Ministry of Education (MoE), Government of India, through the Prime Minister Research Fellowship (PMRF) Program under Grant 2702854. The Associate Editor coordinating the review process was Dr. Adam G. Polak. *(Corresponding author: Udit Satija.)*

The authors are with the Department of Electrical Engineering, Indian Institute of Technology Patna, Patna, Bihar 801106, India (e-mail: arka_2121ee34@iitp.ac.in; udit@iitp.ac.in).

Digital Object Identifier 10.1109/TIM.2023.3292953

six-class classification of respiratory disease (bronchiolitis, bronchiectasis (BRON), pneumonia, upper respiratory tract infection (URTI), COPD, and healthy) using lung sound signals from international conference on biomedical health informatics (ICBHI) 2017 challenge database [4] only.

1) Deep Neural Network-Based Multiclass Respiratory Disease Categorization: Perna [15] had utilized a basic convolutional neural network (CNN) architecture to classify lung sound in three classes: chronic disease, nonchronic disease, and healthy, using MFCC features and achieved the highest accuracy of 82%. Similarly, Pham et al. [1] have used teacher–student mechanism-based CNN-mixture-of-experts (CNN-MoE) architecture, to classify respiratory diseases into three categories. They experimented with different time–frequency images, such as mel spectrogram and gamma tone spectrogram, and achieved a sensitivity of 96% and a specificity of 82% in three-class classification. Nevertheless, when real-time AI-based diagnostic systems are taken into account, this three-class chronic categorization loses its significance, since the doctor or the patient is more interested in learning the exact diagnosis of the sickness, i.e., the specific ailment to which that person is affected. Basu and Rana [14] have introduced a complex 1-D recurrent neural network architecture using gated recurrent units (GRUs) to classify the lung sound signals taken from the ICBHI database into six pathological classes. Their technique has achieved an overall accuracy of 95.67% for six-class respiratory disease classification. Since the ICBHI 2017 database is heavily imbalanced due to the presence of an ample amount of instances in the COPD class, García-Ordás et al. [16] explored the potential of variational autoencoder to augment the minority classes. After augmentation, the training data were fed to a deep CNN model, and in conjunction with this CNN model, an outstanding classification result worth 98.8% sensitivity and 98.6% specificity has been achieved for six-class classification.

2) Lightweight DLM-Based Multiclass Respiratory Disease Categorization: In recent years, biomedical signal analysis solutions have been developed with the idea that the developed algorithms should be deployable in either edge computing devices or in low-resource microcontrollers [2], [17], [18]. The basic drawback of the aforementioned vanilla deep-learning models (DLMs) is their high-computational cost in terms of the number of floating point operations (FLOPs) [18], the number of parameters [18], the amount of storage space required, and high latency, which make them challenging to attain real-time performance with minimal power consumption on resource-limited computing devices [17], [18]. Graphical processing units (GPUs) have been widely employed to speed up the neural network computation, which considerably decreases the computational time by utilizing the parallel processing characteristic. However, GPUs have high power consumption, making them unsuitable for real-time clinical applications. In this scenario, establishing a lightweight DLM for creating a real-time respiratory disease categorization system becomes critical. Shuvo et al. [2] have used bandpass filtering (BPF) and empirical mode decomposition (EMD)-based scale selection followed by a lightweight CNN fed with continuous wavelet transform (CWT)-based scalogram representations to classify lung sounds in six pathological categories. The proposed CNN architecture uses a total of four CNN layers with 64, 64, 96, and 96 filters followed by a flattening layer and six fully connected layers. A total of 3.8 million (M) parameters and 0.86 gillion (G) FLOPs are required to configure the CNN model. By employing this lightweight CNN, an accuracy of 98.7%, a sensitivity of 98.6%, and a specificity of 100.0% have been achieved.

3) Asthmatic Lung Sound Incorporation and Classification: Asthma is one of the minority classes in the ICBHI dataset with only two sound recordings, making deep-learning approaches unreliable for its inclusion in most classification methods used in [2], [14], and [16]. Altan et al. [19] have shown an interesting result on asthma classification by utilizing lung sounds from their own recorded database. Hilbert transform (HHT)-based time- and frequency-domain feature extraction process followed by a deep belief network (DBN) has been utilized, and an accuracy rate of 84.61% has been achieved for binary class classification (asthma versus healthy). To distinguish asthmatic lung sound signals from healthy ones, Tripathy et al. [20] developed an empirical wavelet transform (EWT)-based temporal and spectral feature extraction method, followed by ML classifiers, such as support vector machine (SVM), k-nearest neighbor (KNN), random forest (RF), and light gradient boosting machine (LGBM), using the lung sounds from chest wall lung sound database [21] and achieved an accuracy of 80.35%.

All the aforementioned classification methods use either traditional ML classifiers with handcrafted features and/or CNN-based vanilla deep-learning architectures, which fail to derive the accurate representation of the highly varying time–frequency content of lung sound signals and lead to poor classification performance. In addition to that high memory requirement, computational complexity is also associated with these methods, which are not suitable for creating real-time clinical solutions. Therefore, there is a need to develop a novel lightweight deep-learning network that can provide accurate distinct feature representations from the lung sound signals of different diseased conditions and can achieve a higher classification performance.

B. Objective and Key Contributions

In this article, one of the major objectives is to provide an automated algorithmic approach that can categorize lung sounds in a variety of diseased states. Another objective of this present research work is to propose a lightweight deep-learning architecture that can classify lung sounds accurately while keeping parameter size and computational complexity less. The majority of these respiratory diseases have almost similar kind of symptoms; therefore, it becomes difficult for the doctor to predict the actual disease just by hearing the lung sound only and requires additional tests, such as spirometry test [3], FeNo test, mucus test, and so on. Due to the similar characteristics of lung sounds, even an experienced pulmonologist may make a mistake during an auscultation examination if they do not use other intelligent diagnostic

techniques. In this work, we have exploited the potential of mel spectrogram images for respiratory disease classification using our proposed novel lightweight deep-learning architecture, namely, the respiratory disease lightweight inception network (RDLINet). The proposed framework includes the subsequent stages: 1) preprocessing of the raw lung sound signal; 2) mel spectrogram extraction; and 3) classification using the proposed lightweight inception architecture: RDLINet. For this work, we have extensively used all possible publicly available databases suitable for lung sound-based disease classification tasks, unlike the existing methods that have exploited a single database. In addition, this is the first study to classify respiratory diseases into seven categories using lung sounds by incorporating asthma class into the categorization problem for the first time. The novel contributions of the proposed framework are itemized as follows.

1) Designing a novel lightweight DLM, namely, RDLINet, to classify the lung sound efficiently, while keeping the architecture lightweight in terms of total trainable parameters and model storage size.
2) Classification of seven respiratory diseases for the first time, utilizing three publicly available lung sound databases: ICBHI 2017 challenge database [4], RespiratoryDatabase@TR [3], and chest wall lung sound database [21]. Engaging all the databases also ensures the robustness of the classification mechanism, as the DLM is trained with a wide variety of lung sounds.
3) Computing the ablation study and classification report containing the statistics of layers, parameters, accuracy, precision, recall, $F1$ score, and so on, in order to have a thorough performance/classification accuracy analysis of the proposed lightweight DLM.

II. Description of Database

In this section, we discuss the brief information about three publicly available databases used for this study.

1) ICBHI 2017 Challenge Database (D1): The ICBHI 2017 challenge database, a benchmark data repository of lung auscultation sounds, is used in this study for respiratory disease detection [4]. The database includes 920 audio signals collected from 128 subjects who are either healthy or have one of the following respiratory disorders: asthma, BRON, bronchiolitis, COPD, URTI, LRTI, or pneumonia. The database contains recording worth 5.5 h. The lung sounds are recorded from different auscultation sites: 1) anterior right; 2) anterior left; 3) posterior right; 4) posterior left; 5) lateral right; and 6) trachea. The length of the audio signal is nonuniform, ranging from 10 to 90 s, and the signals are sampled with different sampling frequencies ranging from 4 to 44.1 kHz. The database also contains the information regarding respiratory cycles, which are annotated as normal, crackle, wheeze, and both (contains both wheeze and crackle). The demographic information and data collection protocols are described in [4]. Since LRTI was discovered to have an insufficient amount of segmented samples, these recordings were not taken into consideration for this study.

2) Chest Wall Lung Sound Database (D2): This is one of the largest publicly available lung sound databases on respiratory disease detection using lung sounds [21]. The database includes 336 lung sound recordings from 112 subjects representing either healthy or one of the six pathological cases: asthma, BRON, COPD, pneumonia, heart failure, and pleural effusion. The lung sound signals are captured using a Littmann 3200 digital stethoscope at King Abdullah University Hospital, Jordan, and the signals are sampled at 4 kHz. The audio signal's duration is irregular, varying from 10 to 50 s. To evaluate the efficacy of the proposed technique, we have primarily concentrated on the healthy, asthmatic, BRON, pneumonia, and COPD classes in this study.

3) RespiratoryDatabase@TR (D3): RespiratoryDatabase@TR [3] is a publicly available multimedia database that includes 12-channel lung sound recordings and four-channel heart sound recordings from patients with varying degrees of COPD. The lung sound signals are recorded using a Littmann 3200 digital stethoscope with 4-kHz sampling frequency at Mustafa Kemal University, Turkey. Depending on the lung function test and exacerbation level of the COPD disease, the subjects are labeled as COPD0. COPD1, COPD2, COPD3, or COPD4 with the help of two expert pulmonologists. The recordings of lung sound last at least 17 s. The lung sound signal is acquired at 12 distinct locations from the posterior and anterior sides of the body. The database contains lung sounds of 41 COPD patients (five patients with COPD0, five patients with COPD1, seven patients with COPD2, seven patients with COPD3, and 17 patients with COPD4) with varying degrees of severity, ranging in age from 38 to 68. In this study, we have relabeled all the recordings as COPD class. This will also guarantee the robustness of the trained DLM, as the DLM will be able to recognize the COPD disease efficiently due to the fact that it was trained with all potential COPD variations.

III. Proposed Methodology

The main objective of the current research work is to use a novel lightweight DLM to classify a wide range of respiratory disorders by utilizing lung sounds obtained from a digital stethoscope. The block diagram of the present research is illustrated in Fig. 1, which consists of three major stages: 1) preprocessing of the lung sound signal; 2) time–frequency representation (TFR) extraction for each processed signal; and 3) the mel spectrograms are fed to the lightweight CNN model, namely, RDLINet for the classification of the associated respiratory disease. The individual steps are covered in detail in Sections III-A–III-D.

A. Preprocessing

The preprocessing stage includes four submodules: 1) resampling; 2) temporal snippet generation (split the input signal or time series as a chunk of subseries, called snippets); 3) baseline removal using discrete Fourier transform (DFT)-based filtering; and 4) normalization. Since we are utilizing different databases for respiratory disease classification using lung sounds, it can be observed that lung sounds exhibit a wide

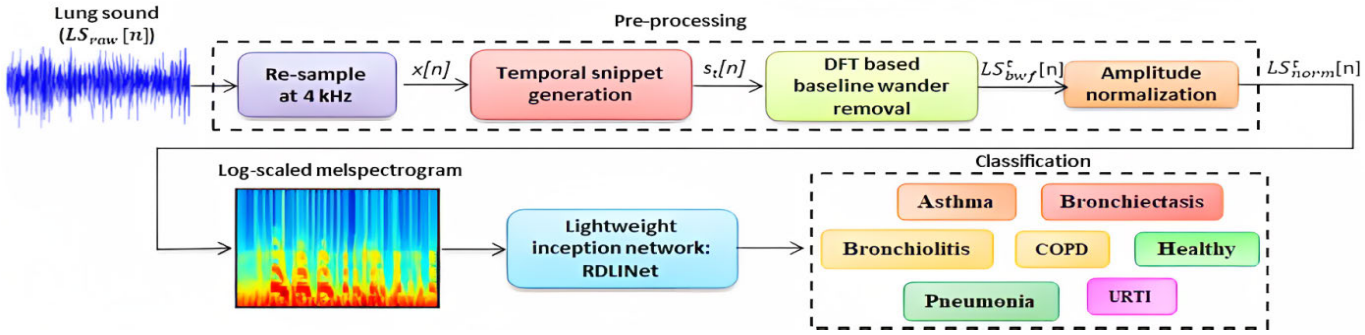

Fig. 1. Block diagram of the proposed RDLINet for seven-class respiratory disease classification.

range of sampling frequency [3], [4], [21]. Therefore, the raw lung sound signals ($\mathrm{LS}_{\mathrm{raw}}[n]$) are resampled to 4 kHz [13]. After resampling, the signals ($\mathrm{LS}_{\mathrm{re}}[n]$) are passed through a temporal snippet generator module, where the whole signal is framed to 5-s window (u) while keeping 50% overlap (r) with next adjacent frame. Say, $\mathrm{LS}_{\mathrm{re}}[n]$ be the input of the snippet generating module, and the output is a set of snippets denoted as $\{s\} = \{s_1[n], s_2[n], \ldots, s_T[n]\}$, where $s_t[n]$ is refereed as a temporal snippet, t denotes the index, and T is the number of snippets generated from $\mathrm{LS}_{\mathrm{re}}[n]$. Mathematically, this process can be expressed as follows:

$$s_t[n] = \mathrm{LS}_{\mathrm{re}}[u \cdot (t-1) \cdot (1-r) + n], \qquad t = 1, 2, \ldots, T. \tag{1}$$

Thereafter, the baseline wandering (BW) component (0–1 Hz) is removed from these individual temporal snippets by using a DFT-based filtering technique [22]. The mathematical foundation of the DFT-based filtering technique is discussed below. Initially, the DFT of the tth temporal snippet ($s_t[n]$) is calculated as follows:

$$S_t[k] = \sum_{n=0}^{N-1} s_t[n] e^{\frac{-j2\pi nk}{N}}. \tag{2}$$

The frequency of a BW component typically spans from 0 to 1 Hz. Therefore, by excluding the frequencies that are below 1 Hz, we can reject the BW part from the DFT coefficient vector. The DFT coefficient index k for the frequency component of f Hz is calculated as $k = \lceil (fN/f_s) \rceil$, where f_s indicates the rate of sampling of the resampled lung sound. Thereafter, the BW-free lung sound snippet ($s^t_{\mathrm{bwf}}[n]$) can be computed by taking inverse DFT of the thresholded DFT vector ($\tilde{S}_t[k]$), as the following:

$$s^t_{\mathrm{bwf}}[n] = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{S}_t[k] e^{\frac{j2\pi nk}{N}} \tag{3}$$

where the thresholded DFT index vector is created as follows: $\tilde{S}_t[k] = [0, \ldots, 0, S_t[k+1], \ldots, S_t[N-k-1], 0, \ldots, 0]$ Finally, the BW-free lung sound snippet's amplitude is normalized in the range of $[-1, 1]$ by using the following equation:

$$s^t_{\mathrm{norm}}[n] = \frac{s^t_{\mathrm{bwf}}[n]}{\max\left|\left(s^t_{\mathrm{bwf}}[n]\right)\right|} \tag{4}$$

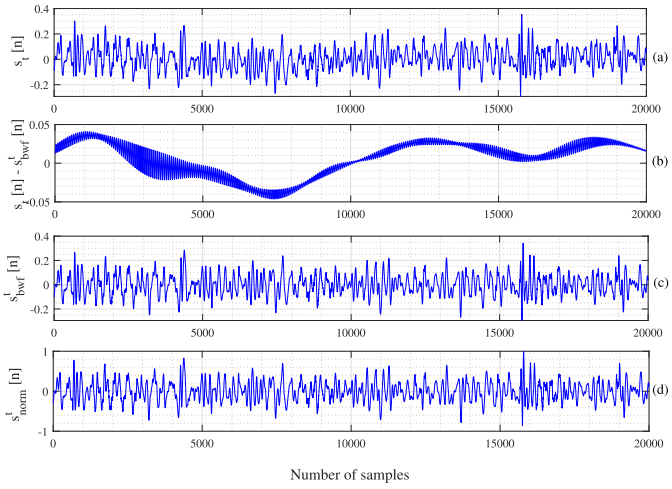

Fig. 2. (a) Raw lung sound signal snippet taken from $D2$ with a BRON-affected individual. (b) Extracted BW component. (c) BW-removed snippet. (d) Normalized lung sound snippet.

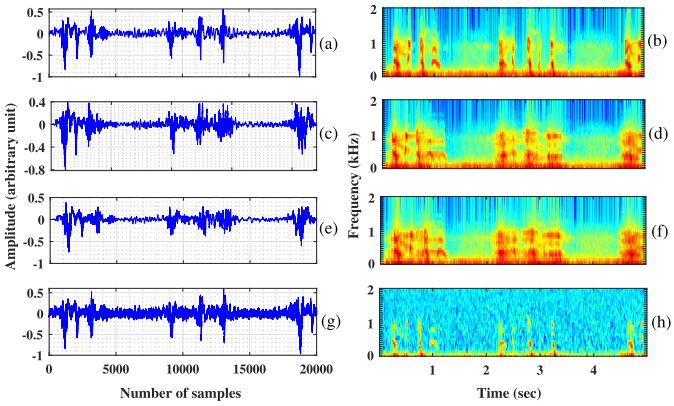

Fig. 3. Time domain and mel spectrogram visualization of (a) and (b) original lung sound signal from a COPD-affected person, (c) and (d) time-stretched augmented version (with a stretching factor of 0.9), (e) and (f) pitch-shifted augmented version (with a pitch-shift factor of −2), and (g) and (h) noise addition-based augmented audio data with an SNR of 10 dB.

where $s^t_{\mathrm{norm}}[n]$ denotes the tth normalized lung sound snippet. Fig. 2(a) illustrates the raw lung sound signal taken from $D2$ with a BRON-affected individual, Fig. 2(b) illustrates the removed BW component, Fig. 2(c) illustrates the BW-removed snippet, and Fig. 2(d) illustrates the normalized lung sound snippet.

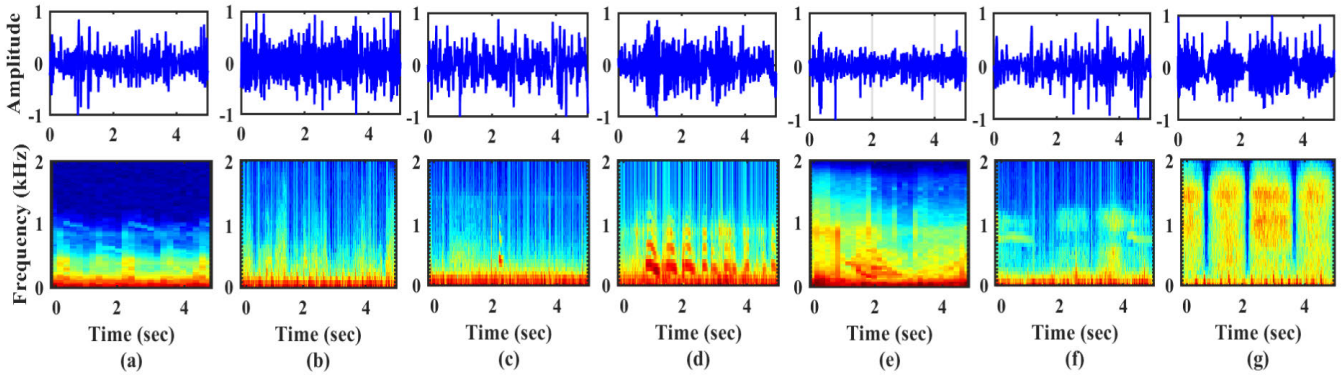

Fig. 4. Time-domain plot (first row) and mel spectrogram (second row) TFR of lung sound signals from different pathological disorders. (a) Healthy. (b) BRON. (c) Bronchiolitis. (d) COPD. (e) Asthma. (f) Pneumonia. (g) URTI.

B. Data Augmentation

We have used three separate databases in this study, as our primary goal is to categorize a wide range of respiratory disorders. However, using all the audio signals from the three databases leads to a problem of class imbalance, where COPD is the majority class and the rest belong to the minority class. To deal with this problem, we have adopted different time-domain audio data augmentation techniques, such as the following.

1) Time Stretching: Another method of lung sound augmentation involves slowing or speeding up audio samples while keeping the pitch constant [23]. In this article, lung sounds from the minority class were augmented by two stretching factors: {0.7, 0.9}.

2) Pitch Shifting: Pitch shifting alters the pitch of the audio signal by either increasing or decreasing the pitch while maintaining the audio signal's length. In [24], the significance of the pitch-shifting process for CNN-based sound categorization is studied. To increase the sample size in minority classes, we have employed two pitch factors of $\{-2, 1\}$.

3) Adding Noise: We have employed white noise addition as another lung sound augmentation technique to increase minority class samples.

Fig. 3(a) and (b) illustrates the temporal and mel spectrogram visualization of the original lung sound signal from a COPD-affected person, Fig. 3(c) and (d) denotes the time domain and mel spectrogram of the time-stretched augmented version of the original signal (with the stretching factor of 0.9), Fig. 3(e) and (f) refers to the pitch-shifted augmented version's time-domain plot and mel spectrogram TFR (with a pitch shift of -2), and Fig. 3(g) and (h) refers to the noise addition-based augmented audio data with an SNR of 10 dB.

C. Extraction of TFR

Lung sound signals include significantly oscillating amplitudes and fluctuating frequency components, making analysis challenging. Therefore, it is necessary to transform the signal from one domain to another in order to gain a comprehensive understanding of them. Transformation techniques enable simultaneous capture of the details of time-domain and frequency-domain signals. In this regard, TFR can capture the spectral variation of the lung sound signal over time. For the lung sound classification problem, mel spectrogram [1], MFCC [1], [14], constant Q transform (CQT) spectrogram [1], and scalogram [2], [10] have been the most widely used input TFR. In this article, we have extracted mel spectrogram TFR from each of the normalized lung sound snippets ($s^t_{\text{norm}}[n]$). To extract the mel spectrogram TFR, initially, a short-time Fourier transform (STFT) (ST_{ls}) from $s^t_{\text{norm}}[n]$ is given as follows:

$$\text{ST}_{\text{ls}}[l, f] = \sum_{n=0}^{N-1} s^t_{\text{norm}}[n] \cdot \mathcal{W}[n - l\mathcal{H}] \cdot e^{-j\frac{2\pi nf}{N}} \tag{5}$$

where $\mathcal{W}[n]$ refers to the Hamming window with 1024 samples, with an overlap length ($\mathcal{H}$) of 512 samples. Thereafter, the hertz frequencies (f) are projected to mel-scale frequency (f_{mel}) for constructing the mel-filter banks. The mel-scale conversion is carried out by employing the following equation [25]:

$$f_{\text{mel}} = 2595 \cdot \log\left(1 + \frac{f}{700}\right). \tag{6}$$

To extract the mel spectrogram we have considered 64 mel filters, as it has also been established in [26], that a larger number of mel filters often degrade the performance of the DLM. To extract the mel spectrogram, the mel filters are multiplied with each of the STFT frames ($\text{ST}_{\text{ls}}[l, f]$) [25]. Finally, a log transform is used on the amplitudes of the mel spectrogram to produce the log-scaled mel spectrogram. Thereafter, these 2-D mel spectrograms are converted to three-channel images by using "jet" color map [25]. Fig. 4(a)–(g) illustrates the time-domain plot (first row) and mel spectrogram (second row) of lung sound signals from different pathological disorders: Fig. 4(a)—healthy, Fig. 4(b)—BRON, Fig. 4(c)—bronchiolitis, Fig. 4(d)—COPD, Fig. 4(e)—asthma, Fig. 4(f)—pneumonia, and Fig. 4(g)—URTI.

D. Proposed Lightweight Inception Network

Different CNN architectures [2], [10], [27] use the TFR image of the lung sound signal, as input has been used in recent years to classify respiratory diseases from pathological lung sounds. However, the fundamental disadvantage of these

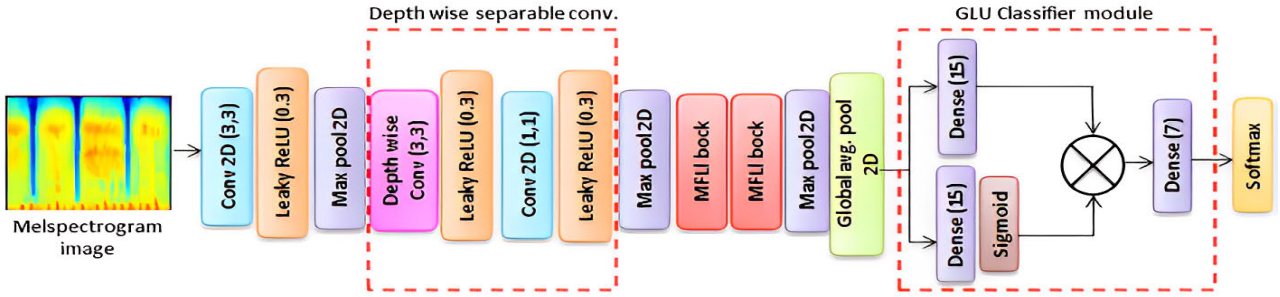

Fig. 5. Detailed architecture of the proposed lightweight inception network: RDLINet.

CNN architectures is their high-computational cost, which limits their ability to be implemented on edge computing devices or embedded processors with limited resources. In addition, cloud computing is rendered obsolete by the rising RAM requirements for computationally intensive training. These factors contribute to the growing popularity of lightweight CNN architectures in the deep-learning research field [28], due to their quick processing speeds and small size without sacrificing accuracy. In this study, we developed a novel lightweight DLM called RDLINet that can diagnose seven distinct respiratory disorders by extracting prominent features from TFR images obtained from the lung sound signal. We drew inspiration from GoogLeNet [29] and MobileNet [30] to design a multiscale filtering lightweight inception (MFLI) block used inside the RDLINet and trained it in an end-to-end fashion using lung sound mel spectrogram data from $D1$–$D3$. Fig. 5 illustrates the architecture of the proposed lightweight inception network. The following proposed principles are used in the architecture design of the RDLINet.

1) The use of depthwise separable convolution (combination of depthwise and pointwise convolution or standard $\{1 \times 1\}$ convolution) reduces the overall computation complexity and parameter size of the CNN model [30]. For example: if the size of the input tensor is $D_G \times D_G \times M$, the size of the filter is $D_F \times D_F$, and the output tensor size is $D_O \times D_O \times N_O$

$$\frac{\text{Parameters depthwise separable conv.}}{\text{Parameter in standard conv.}} = \frac{1}{N_O} + \frac{1}{D_F^2} \tag{7}$$

thereby drastic reduction in total parameter size [30].

2) The majority of the existing CNN models utilizes $\{3 \times 3\}$ filters in the depthwise separable convolution layer. However, the computational expense of such a convolution process is dominated by the pointwise convolution mechanism. For instance, let us say the size of the input tensor is $D_G \times D_G \times M$, and the size of the filter is $D_F \times D_F$; therefore, the total number of operations in the depthwise convolution part is $D_G^2 \times D_F^2 \times M$, while the pointwise convolution on N channels requires $D_G^2 \times M \times N_O$ operation [31]

$$\frac{\text{No. of operation in depthwise conv.}}{\text{No. of operation in pointwise conv.}} = \frac{N_O}{D_F^2}. \tag{8}$$

Therefore, using various sizes of filters, such as $\{3 \times 3\}$ and $\{5 \times 5\}$, in the depthwise component does not result in an increase in computing overhead [31].

3) The vanilla inception block [29] is modified by replacing the standard convolution blocks with depthwise separable convolutions (which were introduced in MobileNet [30]) within the inception block. The modified inception block has four distinct parallel paths to process features from a single input tensor. This aids in the processing of multiscale features and, additionally, reduces the parameter size. The modified inception block is named MFLI block and contains four different paths to process multiscale features from a single input tensor while keeping the parameter size less. Fig. 6 illustrates the proposed architecture of MFLI block.

4) Use of a global average-pooling (GAP) layer instead of a flattened layer-based fully connected layer also helps in reducing the number of trainable parameters [32], [33]. As well, the GAP layer also prevents overfitting for the overall structure by acting as a structural regularizer itself [33].

5) Also, we have used a gated linear unit (GLU)-based classifier, which also helps in efficient classification [34].

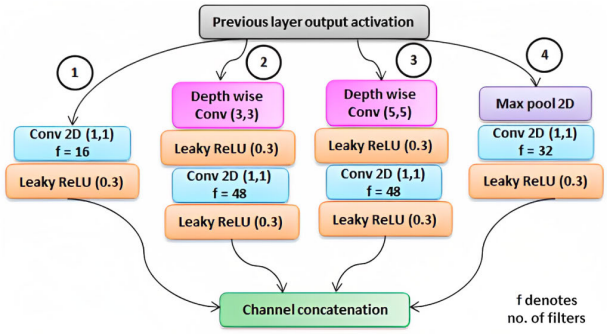

Fig. 6. Proposed MFLI block.

Our proposed RDLINet consists of a total of 43 layers: one input layer, one $\{3 \times 3\}$, four $\{1 \times 1\}$ standard convolutional layer, three $\{3 \times 3\}$ and two $\{5 \times 5\}$ depthwise convolution layer, five pointwise convolution layers, five max-pooling layers, 13 leaky rectified linear unit (ReLU), one sigmoid activation layer, one GAP layer, three dense layers, and one softmax layer. Initially, the 2-D log-scaled mel spectrogram is converted to a three-channel image with a resolution of

$(64 \times 38 \times 3)$ and fed to the standard convolution layer with $\{3 \times 3\}$ kernel, 16 filters, and stride 2 and followed by leaky ReLU activation. Therefore, the resultant dimension becomes $(32 \times 19 \times 16)$. After that, we apply a max-pooling operation with stride 2, which reduces the input tensor's spatial dimension to $(16 \times 10 \times 16)$ for the proceeding layers. Thereafter, this tensor is fed to a depthwise separable convolution block, which comprises 3×3 depthwise convolution block followed by 1×1 pointwise standard convolution block. The output tensor after the depthwise separable convolution has a shape of $(16 \times 10 \times 32)$, which is then fed to leaky ReLU activation followed by max-pooling operation with stride 2 producing a tensor of size $(8 \times 5 \times 32)$. Thereafter, this output is propagated through two MFLI blocks successively. The proposed MFLI block shown in Fig. 6 is a modified version of vanilla inception block [29], which comprises of four parallel paths and receives input from the previous layer's output tensor represented as $\mathscr{F}^x$. This MFLI block enables the model to learn multiscale feature representation from the single input tensor. Path 1 consists of $\{1 \times 1\}$ standard convolution followed by leaky ReLU activation, and paths 2 and 3 comprise of depthwise separable convolutions with $\{3 \times 3\}$ and $\{5 \times 5\}$ convolutions being used in depthwise convolution part, respectively. The mathematical operation of each path of this MFLI block can be represented as follows:

$$\mathscr{F}^{P_1} = \sigma^{\mathrm{LR}}\left(\mathscr{G}^{\mathrm{conv\_(1\times1)}}\left(\mathscr{F}^x\right)\right) \tag{9}$$

$$\mathscr{F}^{P_2} = \sigma^{\mathrm{LR}}\left(\mathscr{G}^{\mathrm{point}}\left(\sigma^{\mathrm{LR}}\left(\mathscr{G}^{\mathrm{depth\_(3\times3)}}\left(\mathscr{F}^x\right)\right)\right)\right) \tag{10}$$

$$\mathscr{F}^{P_2} = \sigma^{\mathrm{LR}}\left(\mathscr{G}^{\mathrm{point}}\left(\sigma^{\mathrm{LR}}\left(\mathscr{G}^{\mathrm{depth\_(5\times5)}}\left(\mathscr{F}^x\right)\right)\right)\right) \tag{11}$$

$$\mathscr{F}^{P_4} = \mathscr{G}^{\mathrm{MP}}\left(\sigma^{\mathrm{LR}}\left(\mathscr{G}^{\mathrm{conv\_(1\times1)}}\left(\mathscr{F}^x\right)\right)\right) \tag{12}$$

where $\mathscr{F}^{P_i}$, $\mathscr{G}^{\mathrm{point}}$, $\mathscr{G}^{\mathrm{depth\_(k\times k)}}$, $\mathscr{G}^{\mathrm{MP}}$, σ^{LR}, and $\mathscr{G}^{\mathrm{conv\_(1\times1)}}$ represent output the of ith parallel path of MFLI block, pointwise convolution, depthwise convolution with $\{k \times k\}$ kernel, max pooling, leaky ReLU activation, and $\{1 \times 1\}$ convolution. Finally, the output of all these paths is concatenated across the channel or depth, which is denoted as $\mathscr{O}^{\mathrm{MFLI}}$ and mathematically described as follows:

$$\mathscr{O}^{\mathrm{MFLI}} = \text{Channel Concat}\left[\mathscr{F}^{P_1}, \mathscr{F}^{P_2}, \mathscr{F}^{P_3}, \mathscr{F}^{P_4}\right]. \tag{13}$$

After propagating the tensor of size $(8 \times 5 \times 32)$, through two MFLI blocks successively, the final output yields a dimension of $(8 \times 5 \times 64)$, which is further downsampled by max-pooling layer to $(4 \times 2 \times 64)$. Thereafter, we have employed a GAP layer, which produces a single dimensional vector $\mathscr{F}^{\mathrm{GAP}} \in \mathscr{R}^{64\times1}$ from the tensor of $(4 \times 2 \times 64)$. Following this, we have employed a GLU-based [34] classifier, which can be mathematically represented as follows:

$$\mathscr{O}^{\mathrm{GLU}} = \mathscr{G}^{D_7}\left\{\mathscr{G}^{D_{15}}\left(\mathscr{F}^{\mathrm{GAP}}\right) \odot \sigma\left(\mathscr{G}^{D_{15}}\left(\mathscr{F}^{\mathrm{GAP}}\right)\right)\right\}. \tag{14}$$

Finally, the output of the GLU module ($\mathscr{O}^{\mathrm{GLU}} \in \mathscr{R}^{7\times1}$) is fed to the softmax activation function [35] to produce class probability values for seven-class respiratory disease classification. Finally, the model is trained in an end-to-end fashion by using the gradient decent-based weight modification method and categorical cross entropy as the loss function for seven-class classification problem. Table I illustrates the total number of parameters required to configure the proposed lightweight RDLINet. From Table I, it can be observed that the model comprises 12 942 parameters to process the input TFR image.

TABLE I
DESCRIPTION OF THE PARAMETER SIZE OF RDLINET

Layer type	Output size	Parameters
Input layer	64*38*3	0
Conv 2D	32*19*16	448
Leaky ReLU	32*19*17	0
Max pool 2D	16*10*16	0
Depthwise conv 2D	16*10*16	160
Leaky ReLU	16*10*16	0
Pointwise conv 2D	16*10*32	544
Leaky ReLU	16*10*32	0
Max pool 2D	8*5*32	0
MFLI block 1	8*5*64	3264
MFLI block 2	8*5*64	6464
Max pool 2D	4*2*64	0
Global average pool 2D	64*1	0
GLU module	7*1	2062
Total		**12942**
FLOPs		**0.000794 G**

IV. RESULTS AND DISCUSSION

In this section, we evaluate the effectiveness of the proposed RDLINet for the classification of respiratory diseases utilizing lung sounds from $D1$–$D3$.

A. Classification Tasks and Performance Metrics

In recent years, various studies have been conducted to classify respiratory diseases using lung sound signals taken from either $D1$ [1], [2], [27] or private volunteer database [19]. However, we have utilized all publicly available databases ($D1$–$D3$) to classify a wide spectrum of respiratory diseases. Thereby, our **Task 1** is to classify seven respiratory diseases utilizing lung sound from the combined database. **Task 2** is to classify six respiratory diseases utilizing lung sounds only from $D1$ using the same RDLINet. **Task 3** is to classify asthma and healthy subjects using lung sounds from both $D1$ and $D2$ databases. In [2] and [27], $D1$ is not used for asthma classification, as it is one of the minority classes; however, as we are using $D2$ also for asthma classification task, we have taken into account both databases. In addition, to the best of our knowledge, this is the third work on asthma classification based on lung sounds following Altan et al. [19]. To evaluate the performance of the proposed RDLINet, we have employed the following performance metrics: accuracy (Acc), specificity (Spe), recall (Rec) or sensitivity (Sen), and precision (Prc) to evaluate our proposed framework similar to [35], [36], and [37]. The performance parameters can be computed from the confusion matrix [38]. The confusion matrix for **Task 1** can be represented as follows:

$$\begin{bmatrix} U_{1,1} & U_{1,2} & \cdots & U_{1,7} \\ U_{2,1} & U_{2,2} & \cdots & U_{2,7} \\ \vdots & \vdots & \ddots & \vdots \\ U_{7,1} & U_{7,2} & \cdots & U_{7,7} \end{bmatrix}$$

where $U_{k,k}$ denotes the diagonal entries of the confusion matrix referring to the correctly classified samples and

TABLE II

MODEL SIMULATION PARAMETERS FOR SEVEN-CLASS RESPIRATORY DISEASE CLASSIFICATION

Simulation parameter	Details
Learning rate	0.008
Optimizer	Adam
Epochs	300
Input image dimension	$64 \times 38 \times 3$
Batch size	128
Loss function	Categorical cross entropy

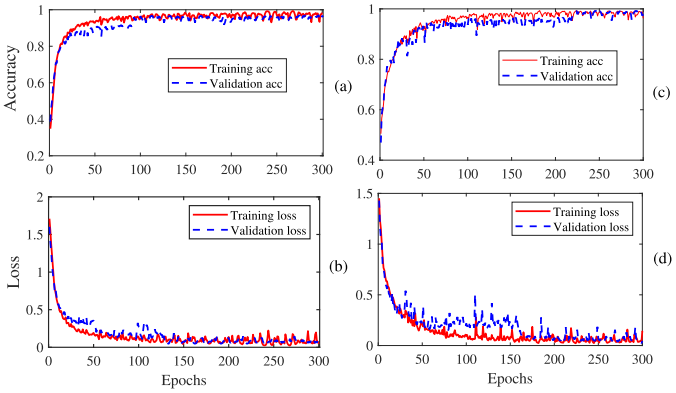

Fig. 7. (a) and (b) Classification accuracy and loss plots of RDLINet for seven-class respiratory disease classification task utilizing all database $D1$–$D3$. (c) and (d) Classification accuracy and loss plots of RDLINet for six-class respiratory disease classification task based on only $D1$.

$U_{k,j}(k \neq j)$ refers to the off-diagonal entries of the confusion matrix, which refer to the samples belonging to the kth class that is misclassified as jth. The accuracy, recall, $F1$ score, and precision of the kth class can be computed as follows [36], [38]: $\text{Acc} = (\text{tp}_k + \text{tn}_k/\text{tp}_k + \text{fp}_k + \text{fn}_k + \text{fn}_k)$, $\text{Rec/Sen} = (\text{tp}_k/\text{tp}_k + \text{fn}_k)$, $\text{Spe} = (\text{tn}_k/\text{fp}_k + \text{tn}_k)$, $\text{Prc} = (\text{tp}_k/\text{tp}_k + \text{fp}_k)$, and $F1$ score $= (2 \times \text{Prc} \times \text{Rec}/\text{Prc} + \text{Rec})$, where tp_k, tn_k, fp_k, and fn_k represent the true positive, true negative, false positive, and false negative of the kth class, which can be computed as follows [38]: $\text{tp}_k = U_{k,k}$, $\text{fn}_k = \sum_{j=1, j\neq k}^{7} U_{j,k}$, $\text{fp}_k = \sum_{j=1, j\neq k}^{7} U_{k,j}$, and $\text{tn}_k = \sum_{j=1}^{7}\sum_{p=1}^{7} U_{j,p} - \text{tp}_k - \text{fp}_k - \text{fn}_k$.

B. Performance Evaluation

The performance of the proposed architecture is evaluated based on training and testing of the network using fivefold cross-validation method for all three classification strategies. First, the whole data are spitted into 80%, 10%, and 10% for training, validation, and testing sets, respectively. Table II illustrates the simulation hyperparameters that have been used to train the proposed lightweight DLM. These hyperparameters are also chosen using the grid search method. Fig. 7(a) and (b) illustrates the training accuracy and loss curves for Task 1; likewise, Fig. 7(c) and (d) illustrates the training accuracy and loss curves for Task 2. From both the training curve sets, it can be observed that our proposed lightweight RDLINet provides a decent amount of classification accuracy for both tasks and does not overfit. To assess the classwise performance and calculate the classification error of our proposed RDLINet, a thorough statistical performance is conducted. The classification report chart presented in Table III shows the classwise

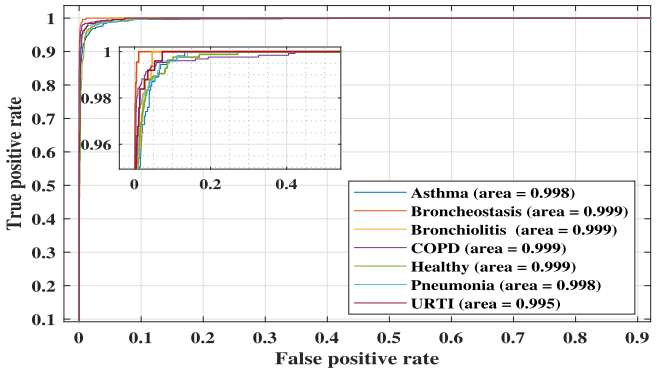

Fig. 8. ROC curve-based classification performance of the RDLINet for all seven classes of Task 1.

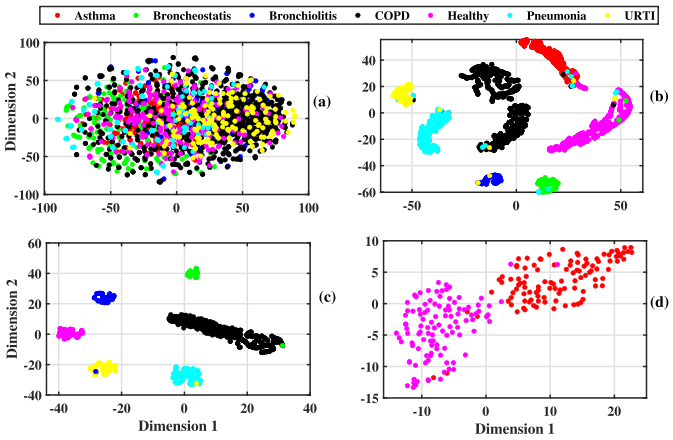

Fig. 9. 2-D t-SNE visualization of (a) raw lung sound signals, and features extracted from the GLU classifier for the classified lung sounds for (b) Task 1, (c) Task 2, and (d) Task 3.

TABLE III

REPORT OF SEVEN-CLASS DISEASE CLASSIFICATION USING RDLINET EVALUATED ON LUNG SOUNDS FROM $D1$–$D3$

Categories	Precision	Recall	F1-score
Asthma	0.94	0.97	0.95
Broncheostasis	0.97	0.97	0.97
Bronchiolitis	0.97	0.93	0.95
COPD	0.99	0.99	0.99
Healthy	0.94	0.97	0.95
Pneumonia	0.96	0.94	0.95
URTI	0.93	0.93	0.93

$F1$ score, recall, and precision for Task 1, demonstrating the overall classification accuracy of the proposed lightweight model for Task 1 to be 96.6%. Fig. 8 illustrates the receiver operating characteristics (ROCs) curve for all classes present in Task 1, and the area under the curve (AUC) values are also mentioned for each class in the same figure. The high AUC values also signify that the proposed model yields high classification performance.

Table IV depicts the foldwise performance of the RDLINet in terms of the aforementioned evaluation metrics for all three classification tasks. It can be observed that Task 1 achieves an average accuracy of 96%, whereas Tasks 2 and 3 achieve an average accuracy of 99.6% and 94.0%, respectively, and

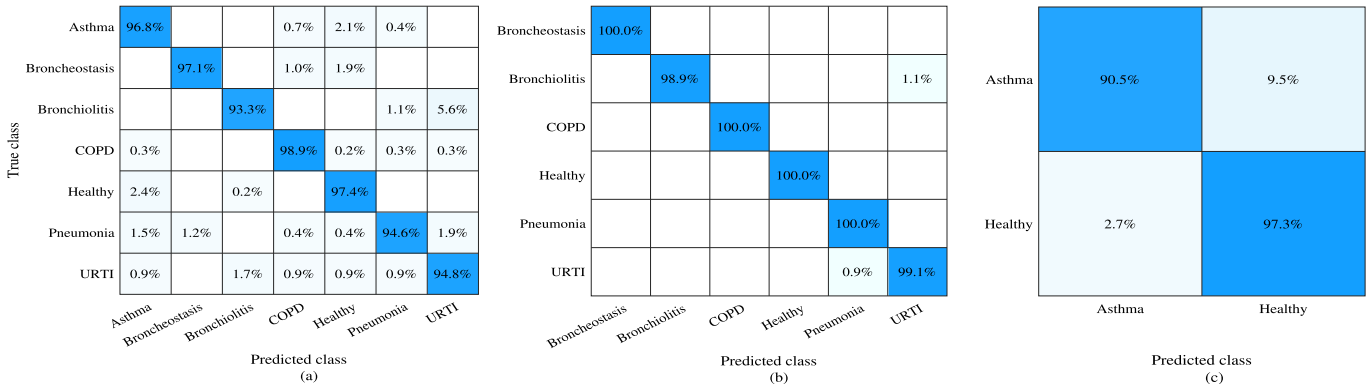

Fig. 10. Normalized confusion matrix for (a) Task 1, (b) Task 2, and (c) Task 3.

TABLE IV
FOLDWISE CLASSIFICATION RESULTS OBTAINED FROM RDLINET

Task no.	Database	Performance metrics (%)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Task 1	D1, D2, D3	Accuracy	96.3	96.8	96.4	97	96.7	96.6
		Recall	95.6	96.1	96.3	95.9	97.1	96.2
		Precision	97.1	95.9	96.2	96.1	96.8	96.4
Task 2	D1	Accuracy	99.8	99.5	99.6	99.7	99.4	99.6
		Recall	100	99.6	100	100	99.8	99.8
		Precision	99.5	99.6	99.4	99.8	99.5	99.5
Task 3	D1, D2	Accuracy	93.7	94.5	93.6	94.1	94.2	94.0
		Recall	90.3	90.4	91.1	90.3	90.2	90.5
		Precision	97.3	96.5	96.6	97	97.1	96.9

outperforms all the published research works on respiratory disease classification.

Fig. 9(a) shows the 2-D t-distributed stochastic neighbor embedding (t-SNE) feature visualization [39] of the raw lung sounds. From the t-SNE plot shown in Fig. 9(a), we clearly observe that all the lung sounds from different classes were randomly scattered in the 2-D plane initially. However, with the application of RDLINet, all the lung sounds from different classes have formed distinct clusters in Fig. 9(b)–(d) for all three tasks. Therefore, the experimental result shown in Fig. 9(b)–(d) indicates that our proposed framework can classify different lung sounds efficiently, by extracting class-dependent features, which leads to the formation of distinct clusters in the 2-D t-SNE plane. To present the supremacy of the RDLINet, we have also shown the normalized confusion matrix for all three classification tasks in Fig. 10. We can observe from Fig. 10 that employing the RDLINet has allowed us to attain high true positive values for each class in each of the three tasks, which in a manner demonstrates the superiority of the framework over other existing frameworks [1], [2], [14], [19], [20], in terms of the classification performance.

DLMs work as a black box in the sense that their learned features lack interpretability and decision-making is opaque. In this article, to nullify this black box nature of the DLM, additionally, we have analyzed the interpretability of our proposed RDLINet by employing a Grad-cam-based network attention visualization technique [40]. Our suggested RDLINet obtains latent representations from input mel spectrogram images by passing them through a sequence of convolution blocks and then making a decision based on these representations. The Grad-cam method is applied on each of the input mel spectrogram images extracted from the test set signals to visualize the region of the mel spectrogram where the model pays closer attention before taking the final decision. In Fig. 11, the first and third rows represent the mel spectrogram image of the lung sound signals of different classes; similarly, the second and fourth rows of Fig. 11 illustrate the Grad-cam visualization based on the corresponding mel spectrogram images. For example, in Fig. 11(b) and (d), we can identify that our proposed RDLINet scrutinizes more on the lower regions of the mel spectrogram images homogeneously in case of healthy lung sound signals, which refers to the low-frequency information present in the actual mel spectrogram representation. Similarly, for the COPD class, we can observe that the network pays more attention to the wheeze events present in the mel spectrogram images in Fig. 11(n) and (p).

C. Sensitivity Analysis With Respect to Different Influencing Parameters

In this section, we explore the significance of several influencing aspects that affect the performance of the proposed framework.

1) Effect of Processing Length of Lung Sound Signal: Since lung sound signals are highly nonstationary signals, we generally process the TFR of the lung sound signal. However, the information of the lung sound time series is present in a sequential manner, and the temporal resolution or processing length of the input signal also influences the classification performance of the DLM. To evaluate the impact of processing length, we have considered 5-, 7-, and 10-s frames of lung sound signal. Fig. 12(a) illustrates the variation in the accuracy of RDLINet with respect to different input lengths of lung sound signal for all three tasks. It can be observed that for 5-s input length, we have achieved the highest accuracy for all three tasks; however, with an increase in processing length, the performance degrades, and for 10-s length, accuracy is drastically reduced.

2) Choice of Proper TFR: In order to extract a compact amount of information from the lung sound signals, which are largely nonstationary in nature, it is advised to transform the time-domain signal into another domain [1]. In such

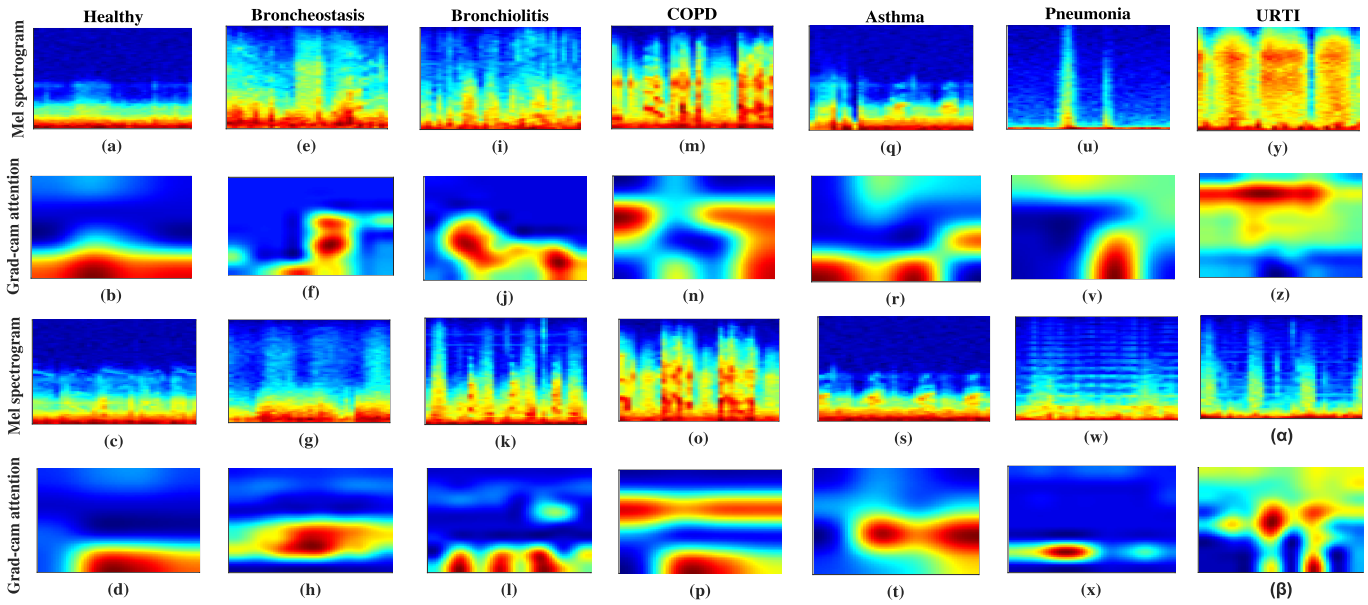

Fig. 11. Mel spectrogram image representation of two different subjects belonging to (a) and (c) healthy, (e) and (g) BRON, (i) and (k) bronchiolitis, (m) and (o) COPD, (q) and (s) asthma, (u) and (w) pneumonia, and (y) and (α) URTI class present in the test set. Corresponding Grad-cam network attention result for (b) and (d) healthy, (f) and (h) BRON, (j) and (l) bronchiolitis, (n) and (p) COPD, (r) and (t) asthma, (v) and (x) pneumonia, and (z) and (β) URTI class based on the fed input mel spectrogram.

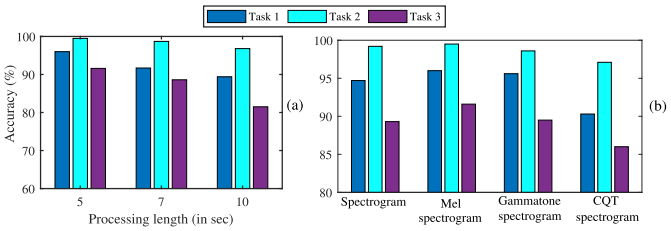

Fig. 12. Effect of using (a) different processing lengths and (b) TFRs on the overall classification accuracy for Tasks 1–3.

a context, TFRs are able to locate the information on the variation of spectral content in respiratory sounds with respect to time. Therefore, the TFR selection may potentially have an impact on the DLM's accuracy. A comparison result analysis is presented in Fig. 12(b), which demonstrates the impact on classification accuracy with regards to various TFRs. We have primarily focused on the mel spectrogram representation, since it has previously been proven to be one of the finest TFRs for the task of classifying audio signals [43]. As seen in Fig. 12(b), the use of mel spectrograms results in the highest classification accuracy for all classification tasks. The classification accuracy is somewhat reduced when using gammatone and simple vanilla spectrograms. Using simple spectrogram, a classification accuracy of 94.7%, 99.2%, and 89.3% and, using gammatone spectrogram, a classification accuracy of 95.6%, 98.6%, and 89.5% are achieved for Tasks 1–3, respectively. However, the CQT spectrogram falls short of the other TFRs in the categorization of lung sounds while yielding a classification accuracy of 90.3%, 97.1%, and 86.0% for Tasks 1–3, respectively.

3) Choosing Optimal Number of MFLI Blocks: In the proposed lightweight architecture, we have used the MFLI block, which facilitates the model to learn multiscale feature representation from the single input tensor; i.e., it provides the flavor of multiscale feature processing. However, choosing the optimal number of MFLI blocks is also a crucial matter. Hence, to determine the optimum number of MFLI blocks, we have provided an ablation study in Table V.

TABLE V

ABLATION STUDY ON CHOOSING OPTIMAL NUMBER OF MFLI BLOCKS

No. of MFMI blocks	Layers	No. of parameters	Acc (%)		
			Task 1	Task 2	Task 3
1	29	6,478	90.3	95.8	87.3
2	**43**	**12,942**	**96.6**	**99.6**	**94.0**
3	55	19,406	91.5	94.6	92.8
4	68	25,870	91.6	93	89.4

TABLE VI

PERFORMANCE COMPARISON OF DIFFERENT MODELS WITH RESPECT TO MODEL EVALUATION METRICS

Model	Model evaluation parameter					
	No. of parameters (M)	Model size	FLOPs	Accuracy (%)		
				Task 1	Task 2	Task 3
VGG-16 [41]	138	1.5GB	15.59361G	89.3	97.6	89.3
MobileNet [30]	4.2	46.9MB	0.57279G	92.3	98.6	90.1
ShuffleNetV2 [42]	5.4	49MB	0.49051G	91.4	98.2	88.5
Custom CNN [2]	3.7	44.85MB	0.86838G	92.8	98.7	89.8
RNN [14]	0.32	3.4MB	0.00492G	90.5	95.6	85.9
RDLINet	**0.0129**	**498KB**	**0.00079G**	**96.6**	**99.6**	**94.0**

From Table V, it can be observed that the highest classification accuracy is achieved with two MFLI blocks, and further increase of MFLI block does not increase the accuracy of the proposed model and, however, leads to the increase in the parameter size of the DLM. Therefore, we have finalized our proposed lightweight RDLINet with two MFLI blocks.

D. Performance Comparison

1) Comparative Analysis in Terms of Trainable Parameter and Computational Complexity Reduction: To prove the lightweight nature of the proposed RDLINet, a thorough comparison study has been provided in Table VI in terms of model size, number of parameters, and achieved accuracy for

TABLE VII
QUALITATIVE PERFORMANCE COMPARISON OF THE PROPOSED RDLINET WITH NOTEWORTHY EXISTING WORKS

Sl. no.	Authors	Pre-processing	Features	Classification method	Task no.	Database	Results		
							Acc(%)	Sen (%)	Spe(%)
1	Basu et al. [14]	Segment into 10 sec frame	MFCC	GRU	2	D1	95.7	95.7	–
2	Shuvo et al. [2]	Framed to 5sec window, BPF (50-2500 Hz), EMD	CWT-scalogram	Lightweight CNN	2	D1	98.7	98.96	100.0
3	Garcia et al. [16]	Segment framing, normalization	Mel spectrogram, variational autoencoder	CNN	2	D1	99.0	98.8	98.6
4	Tripathy et al. [20]	5sec segment frame, normalization, EWT	Peak frequency, peak amplitude, shannon entropy	KNN, SVM, RF, LGBM	3	D2	80.3	84.8	75.2
5	Altan et al. [19]	Frame into 15 sec, 1st order HPF, HHT decomposition	Statistical features from modes	DBN	3	Own	84.6	85.8	77.1
6	**Proposed work**	**Temporal snippet generation, BW removal and normalization**	**Mel spectrogram**	**Lightweight RDLINet**	**1**	**D1, D2, D3**	**96.6**	**96.2**	**98.0**
					2	**D1**	**99.6**	**99.8**	**100**
					3	**D1, D2**	**94.0**	**90.5**	**97.32**

Tasks 1–3. In addition, we have evaluated the computational complexity of the RDLINet by calculating the total number of required multiplication–addition [42] or the number of FLOPs [42] and compared the same with the state-of-the-art (SOTA) deep neural networks, such as visual geometry group-16 (VGG-16) [41], and other lightweight networks, such as MobileNet [30], ShuffleNetV2 [42], and custom lightweight CNN proposed by Shuvo et al. [2]. From Table VI, it can be observed that we have outperformed the SOTA lightweight networks in terms of accuracy while drastically reducing the size of the parameter sets and computational complexity (FLOPs), and consuming less computing and storage resources. In particular, the amount of FLOPs needed for our RDLINet is just 0.00079G, which is around 1099 times less than the total number of FLOPs needed for the custom lightweight CNN proposed by Shuvo et al. [2]. It can be seen that these experimental findings, thus, support the claim that our suggested model is lightweight and suitable for real-time implementation on edge computing hardware or microprocessors with limited resources.

2) Comparative Analysis With the Existing Respiratory Disease Classification Techniques: In this section, the comparative performance of our proposed framework is analyzed with respect to other existing research works on respiratory disease classification based on the lung sounds. Table VII shows the comparative results for all three classification tasks. To date, researchers have only used either $D1$ for six-class classification [2], [14], [16] using lung sound signal. However, in this work, we have used all three databases to train our model, which also ensures the robustness of the model, as we have incorporated all possible pathological lung sounds, which are publicly available. To the best of our knowledge, this is the first work on seven-class respiratory disease detection using lung sounds. From Table VII, it is clear that the RDLINet has outperformed the existing research works. For Task 2, though [2] provides a decent amount of accuracy, the overall algorithm including the preprocessing steps is heavily computationally expensive, since it includes EMD, which is computationally heavy due to iterative decomposition of multicomponent signal [44], thereby increasing the preprocessing time and the same has been discussed by Shuvo et al. [2]. Likewise, for Task 3, we have outperformed the existing works of Altan et al. [19] and Tripathy et al. [20] by a margin of 9% and 13%, respectively. In addition, we have reduced the burden of the handcrafted feature extraction process proposed by Altan et al. [19] and Tripathy et al. [20] by employing the proposed framework.

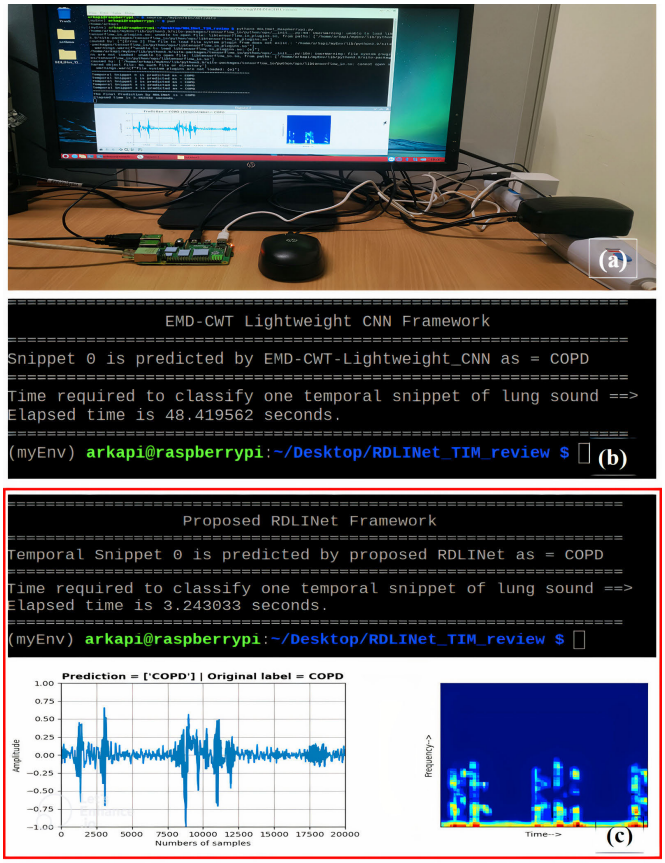

Fig. 13. On-device computational performance evaluation of the proposed RDLINet. (a) Raspberry-Pi 4 hardware setup connected to LCD monitor. Screenshot of the classification result summary of a COPD-affected lung sound signal taken from the ICBHI database using (b) EMD-CWT-based lightweight CNN framework proposed by Shuvo et al. [2] and (c) RDLINet.

V. ON-DEVICE IMPLEMENTATION

The proposed RDLINet-based respiratory disease classification was initially developed in the Google Colab Python environment. After completion of the training of the model, the saved weights have been dumped inside the Rapberry-Pi 4 microcontroller (64-bit quad-core ARM Cortex-A7,

clock frequency 1.5 GHz, Broadcom BCM2711 SoC, 8-GB onboard RAM). The developed algorithm is then transferred to the Thonny IDE, and all the required libraries, such as Librosa, Keras, TensorFlow, and so on, have been installed in the controller. Fig. 13(a) illustrates the on-device hardware implementation layout, and Fig. 13(b) and (c) shows the classification result and the computational time required to classify a single temporal snippet of a COPD-affected lung sound signal taken from the ICBHI database using the EMD-CWT-driven lightweight CNN framework proposed by Shuvo et al. [2] and the RDLINet, respectively. The average time required to classify a single snippet of lung sound using RDLINet is 3.22 ± 0.28 s. Whereas, it takes around 48.53 ± 0.41 s to classify the same single snippet by employing the framework proposed by Shuvo et al. [2], using the same hardware setup of Raspberry-Pi, where the main time-consuming part is the EMD-based CWT extraction process, which introduces high latency to their framework. Therefore, the experimental outcome shows that the suggested framework can be used in a real-time environment, as it maintains a very low latency in comparison with SOTA alternatives on respiratory disease classification by utilizing the lung sound signals.

VI. Conclusion

In this work, we have exploited the potential of mel spectrogram TFR-based respiratory disease classification using a novel lightweight RDLINet. By utilizing the proposed framework, we have achieved an accuracy of 96.6%, 99.6%, and 94.0% for seven-class, six-class, and binary (healthy versus asthma) respiratory disease classification using lung sounds from all the publicly available databases, which outperforms all other research works. In addition, the computational complexity of the RDLINet is compared with a number of well-known DLMs, which support the lightweight nature of the proposed DLM. RDLINet also achieves outstanding classification accuracy in all three classification tasks while keeping the model architecture lightweight in terms of parameter size and number of required operations. Thereby, we think that these traits can facilitate the development of on-device automated respiratory disease categorization systems for real-time clinical use.

Acknowledgment

The authors would like to thank the anonymous reviewers for improving their manuscript in terms of accuracy improvement and computational complexity reduction based on their suggestions.

References

[1] 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 J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 2938–2947, Aug. 2021.

[2] 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 J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2595–2603, Jul. 2021.

[3] G. Altan, Y. Kutlu, Y. Garbi, A. Ö. Pekmezci, and S. Nural, "Multimedia respiratory database (RespiratoryDatabase@TR): Auscultation sounds and chest X-rays," *Natural Eng. Sci.*, vol. 2, no. 3, pp. 59–72, Oct. 2017.

[4] B. Rocha et al., "A respiratory sound database for the development of automated classification," in *Proc. Int. Conf. Biomed. Health Informat.* Singapore: Springer, 2017, pp. 33–37.

[5] A. Rao, E. Huynh, T. J. Royston, A. Kornblith, and S. Roy, "Acoustic methods for pulmonary diagnosis," *IEEE Rev. Biomed. Eng.*, vol. 12, pp. 221–239, 2019.

[6] A. A. Cruz, *Global Surveillance, Prevention and Control of Chronic Respiratory Diseases: A Comprehensive Approach*. Geneva, Switzerland: World Health Organization, 2007.

[7] B. Roy, A. Roy, J. K. Chandra, and R. Gupta, "I-PRExT: Photoplethysmography derived respiration signal extraction and respiratory rate tracking using neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[8] S. Vadrevu and M. S. Manikandan, "A robust pulse onset and peak detection method for automated PPG signal analysis system," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 3, pp. 807–817, Mar. 2019.

[9] S. Chaudhary, S. Sadbhawna, V. Jakhetiya, B. N. Subudhi, U. Baid, and S. C. Guntuku, "Detecting COVID-19 and community acquired pneumonia using chest CT scan images with deep learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8583–8587.

[10] S. Jayalakshmy and G. F. Sudha, "Scalogram based prediction model for respiratory disorders using optimized convolutional neural networks," *Artif. Intell. Med.*, vol. 103, Mar. 2020, Art. no. 101809.

[11] F. Jin, F. Sattar, and D. Y. T. Goh, "New approaches for spectro-temporal feature extraction with applications to respiratory sound classification," *Neurocomputing*, vol. 123, pp. 362–371, Jan. 2014.

[12] T. Fernando, S. Sridharan, S. Denman, H. Ghaemmaghami, and C. Fookes, "Robust and interpretable temporal convolution network for event detection in lung sound recordings," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 7, pp. 2898–2908, Jul. 2022.

[13] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 3, pp. 535–544, Jun. 2020.

[14] V. Basu and S. Rana, "Respiratory diseases recognition through respiratory sound with the help of deep neural network," in *Proc. 4th Int. Conf. Comput. Intell. Netw. (CINE)*, Feb. 2020, pp. 1–6.

[15] D. Perna, "Convolutional neural networks learning from respiratory data," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 2109–2113.

[16] M. T. García-Ordás, J. A. Benítez-Andrades, I. García-Rodríguez, C. Benavides, and H. Alaiz-Moretón, "Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data," *Sensors*, vol. 20, no. 4, p. 1214, Feb. 2020.

[17] M. Saini, U. Satija, and M. D. Upadhayay, "DSCNN-CAU: Deep-learning-based mental activity classification for IoT implementation toward portable BCI," *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8944–8957, May 2023.

[18] G. Sivapalan, K. K. Nundy, S. Dev, B. Cardiff, and D. John, "ANNet: A lightweight neural network for ECG anomaly detection in IoT edge sensors," *IEEE Trans. Biomed. Circuits Syst.*, vol. 16, no. 1, pp. 24–35, Feb. 2022.

[19] G. Altan, Y. Kutlu, A. Ö. Pekmezci, and S. Nural, "The diagnosis of asthma using Hilbert–Huang transform and deep learning on lung sounds," 2021, *arXiv:2101.08288*.

[20] R. K. Tripathy, S. Dash, A. Rath, G. Panda, and R. B. Pachori, "Automated detection of pulmonary diseases from lung sound signals using fixed-boundary-based empirical wavelet transform," *IEEE Sensors Lett.*, vol. 6, no. 5, pp. 1–4, May 2022.

[21] M. Fraiwan, L. Fraiwan, B. Khassawneh, and A. Ibnian, "A dataset of lung sounds recorded from the chest wall using an electronic stethoscope," *Data Brief*, vol. 35, Apr. 2021, Art. no. 106913.

[22] U. Satija, B. Ramkumar, and M. S. Manikandan, "Real-time signal quality-aware ECG telemetry system for IoT-based health care monitoring," *IEEE Internet Things J.*, vol. 4, no. 3, pp. 815–823, Jun. 2017.

[23] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Appl. Sci.*, vol. 6, no. 2, p. 57, Feb. 2016.

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

[25] I. Ozer, "Pseudo-colored rate map representation for speech emotion recognition," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102502.

[26] Z. Zhao et al., "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515–97525, 2019.

[27] 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. 19th Int. Conf. Control, Autom. Syst. (ICCAS)*, Oct. 2019, pp. 804–807.

[28] Q. Yu and L. Sun, "LPClass: Lightweight personalized sensor data classification in computational social systems," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 6, pp. 1660–1670, Dec. 2022.

[29] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[30] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[31] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "BlazeFace: Sub-millisecond neural face detection on mobile GPUs," 2019, *arXiv:1907.05047*.

[32] M. Chakraborty, S. V. Dhavale, and J. Ingole, "Corona-nidaan: Lightweight deep convolutional neural network for chest X-ray based COVID-19 infection detection," *Int. J. Speech Technol.*, vol. 51, no. 5, pp. 3026–3043, May 2021.

[33] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.

[34] N. Shazeer, "GLU variants improve transformer," 2020, *arXiv:2002.05202*.

[35] M. Saini, U. Satija, and M. D. Upadhayay, "One-dimensional convolutional neural network architecture for classification of mental tasks from electroencephalogram," *Biomed. Signal Process. Control*, vol. 74, Apr. 2022, Art. no. 103494.

[36] E. Prabhakararao and S. Dandapat, "Multi-scale convolutional neural network ensemble for multi-class arrhythmia classification," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 3802–3812, Aug. 2022.

[37] A. Shankar, S. Dandapat, and S. Barma, "Seizure types classification by generating input images with in-depth features from decomposed EEG signals for deep learning pipeline," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 4903–4912, Oct. 2022.

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

[39] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

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

[42] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet v2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

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

[44] J. Chen, H. Sun, and B. Xu, "Improvement of empirical mode decomposition based on correlation analysis," *Social Netw. Appl. Sci.*, vol. 1, no. 9, pp. 1–16, Sep. 2019.

Arka Roy (Graduate Student Member, IEEE) received the B.Tech. degree from the Ramkrishna Mahato Government Engineering College, Purulia, India, in 2021. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, Indian Institute of Technology (IIT) Patna, Patna, India.

He is a Prime Minister Research Fellow with IIT Patna. His research interests include biomedical signal processing, deep learning, and graph signal processing.

Udit Satija (Senior Member, IEEE) received the B.Tech. degree in ECE from Rajasthan Technical University (RTU), Kota, India, in 2010, the M.Tech. degree in ECE from The LNM Institute of Information Technology (LNMIIT), Jaipur, India, in 2013, and the Ph.D. degree in cardiovascular signal processing from the Indian Institute of Technology (IIT) Bhubaneswar, Bhubaneswar, India, in 2018.

He is currently an Assistant Professor with the Department of EE, IIT Patna, Patna, India. His research interests include biomedical signal processing and machine learning.