A Lightweight CNN Model for Detecting Respiratory Diseases from Lung Auscultation Sounds using EMD-CWT-based Hybrid Scalogram

Samiul Based Shuvo1, Shams Nafisa Ali1, Soham Irtiza Swapnil1, Taufiq Hasan1, Member, IEEE and Mohammed Imamul Hassan Bhuiyan2, Member, IEEE

Abstract—Listening to lung sounds through auscultation is vital in examining the respiratory system for abnormalities. Automated analysis of lung auscultation sounds can be beneficial to the health systems in low-resource settings where there is a lack of skilled physicians. In this work, we propose a lightweight convolutional neural network (CNN) architecture to classify respiratory diseases using hybrid scalogram-based features of lung sounds. The hybrid scalogram features utilize the empirical mode decomposition (EMD) and continuous wavelet transform (CWT). The proposed scheme’s performance is studied using a patient independent train-validation set from the publicly available ICBHI 2017 lung sound dataset. Employing the proposed framework, weighted accuracy scores of 99.20% for ternary chronic classification and 99.05% for six-class pathological classification are achieved, which outperform well-known and much larger VGG16 in terms of accuracy by 0.52% and 1.77% respectively. The proposed CNN model also outperforms other contemporary lightweight models while being computationally comparable.

Index Terms—Lung auscultation sound, respiratory disease detection, lightweight convolutional neural networks, empirical mode decomposition, continuous wavelet transform, scalogram.

I. INTRODUCTION

UNG diseases are the third largest cause of death in the world [1]. According to the World Health Organization (WHO), the five major respiratory diseases [2], namely chronic obstructive pulmonary disease (COPD), tuberculosis, acute lower respiratory tract infection (LRTI), asthma, and lung cancer, cause the death of more than 3 million people each year worldwide [3]. These respiratory diseases severely affect the overall healthcare system and adversely affect the lives of the general population. Prevention, early diagnosis and treatment are considered key factors for limiting the negative impact of these deadly diseases.

Auscultation of the lung using a stethoscope is the traditional diagnostic method used by specialists and general practitioners for the initial investigation of the respiratory system. Although physicians use various other investigation strategies such as plethysmography, spirometry, and arterial blood gas analysis, lung sound auscultation remains as a vital tool for physicians due to its simplicity and low-cost [5]. The primary classification of these non-periodic and non-stationary sounds consists of two groups: normal (vesicular) and abnormal (adventitious) [6]. The first group is observed when there are no respiratory diseases, while the latter group indicates complications in the lungs or airways [7]. Crackles, wheeze, rhonchus, squawk, stridor, and pleural rub are the commonly known abnormal lung sounds. These anomalies can be differentiated from the normal lung sounds on the basis of frequency, pitch, energy, intensity, timbre, and musicality [8], [9]. Therefore, lung sounds are of particular importance for recognizing specific respiratory diseases and assessing its chronic-nonchronic characteristics. However, the subtle differences between some of the adventitious lung sound classes can be a strenuous task even for a specialist and may introduce subjectivity in the diagnostic interpretation [10]. In this scenario, artificial intelligence (AI)-powered algorithms can be of benefit in automatically interpreting lung sounds, especially in underdeveloped regions of the world, with a scarcity of skilled physicians.

In the past decade, a number of research approaches have been considered and evaluated for automatic identification of respiratory anomalies from lung auscultations sounds. Numerous feature extraction techniques including statistical features [11], entropy-based features [12], wavelet coefficients [13], Mel Frequency Cepstral Coefficients (MFCC) [10], spectrograms [14], scalograms [15] etc. have been adopted in conjunction with a diverse set of machine learning (ML) algorithms [10]–[24].

With the advent of deep learning (DL), new developments have been made in recent times, demonstrating highly promising results in diversified applications, including biomedical engineering and clinical diagnostics [25]–[30]. With the ability of automatic feature learning, deep learning (DL) approaches are more generic and can mitigate the limitations of traditional ML-based methods. In the same vein, DL-based paradigms that are employed in the recent years for the identification of respiratory anomalies and pathologies from lung auscultation data have exhibited highly promising performance [5], [31]–[42]. However, for attaining proper functionality, the deep networks require to undergo an extensive training scheme with a large training dataset that subsequently calls for a considerable amount of time and the engagement of powerful computational resources. As a result, it becomes quite chal-
A graphical overview of the proposed framework. After several generic preprocessing steps, the lung sound signals are converted into scalograms using both conventional and hybrid approaches. The resulting images are further augmented and fed into the proposed lightweight CNN model to carry out a two-way classification of respiratory diseases: (i) Chronic and (ii) Pathological.

Fig. 1. A graphical overview of the proposed framework.
from the spectrograms, recent work has achieved a specificity of 83% and a sensitivity of 96% in ternary respiratory disease classification [40]. For binary classification, the same work has reported specificity and sensitivity of 83% and 99%, respectively. As an extension of [40], a separate study involving the robust Teacher-Student learning schemes with knowledge distillation has been conducted, which resulted in a substantially reduced specificity while maintaining the sensitivity [41].

Since the existing heavily imbalanced datasets of lung auscultations further exacerbate the task of respiratory disease classification, a contemporary study has dealt with this issue by experimenting with several data augmentation techniques, such as SMOTE, Adaptive Synthetic Sampling Method (ADASYN) and Variational autoencoder (VAE) [42]. Among the methods, the VAE-based Mel-spectrogram augmentation strategy, in conjunction with a CNN model, has achieved the best results with 98.5% sensitivity and 99.0% specificity in ternary chronic classification. The strategy has also exhibited an equally sophisticated performance with 98.8% sensitivity and 98.6% specificity in the case of six class respiratory disease classification [42].

Although the scope of DL-based frameworks with a spectrogram-based feature extraction strategy has been investigated in several works for direct classification of respiratory diseases from lung auscultations [40]–[42], to the best of the knowledge of the authors, scalogram based approaches have not been explored in this domain. Additionally, no dedicated lightweight, efficient CNN framework has been developed and investigated for the respiratory disease classification task. Furthermore, none of the studies consider the issue of inpatient dependency in the train-validation split. Inspired by all of these factors, a scalogram based approach in conjunction with a lightweight CNN is proposed in this paper for the prediction of respiratory diseases from lung auscultations, maintaining patient independence. The proposed framework is schematically represented in Fig. 1.

III. MATERIALS AND METHODS

A. ICBHI 2017 Dataset

ICBHI (International Conference on Biomedical Health Informatics) 2017 database is a publicly available benchmark dataset of lung auscultations [47]. It is collected by two independent research teams of Portugal and Greece. The dataset contains 5.5 hours of audio recordings sampled at different frequencies (4 kHz, 10 kHz, and 44.1 kHz), ranging from 10s to 90s, in 920 audio samples of 126 subjects from different anatomical positions with heterogeneous equipment [53].

The samples are professionally annotated considering two schemes: 1. according to the corresponding patients pathological condition, i.e. healthy and seven distinct disease classes, namely Pneumonia, Bronchiectasis, COPD, URTI, LRTI, Bronchiolitis, Asthma and 2. according to the presence of respiratory anomalies, i.e. crackles and wheezes in each respiratory cycle. Further details about the dataset and data collection methods can be found in [53].

B. Data Prepossessing

1) Noise filtering: Since 50 Hz to 2500 Hz is the acknowledged frequency range of the lung auscultation signals [7], the recorded audio signals are filtered with a 6th order Butterworth bandpass filter, thus retaining 50 Hz to 2500 Hz frequency components. Subsequently, all the sample signals are resampled to 22050 Hz for ensuring consistency and normalized to the range [-1,1] for attaining device homogeneity.

2) Segmentation of the sound data: Each of the audio recordings is segmented according to the annotated respiratory cycle timing with a 6s duration each. Samples with a minimum respiratory cycle duration of 3s are taken into account to obtain useful respiratory sound information [40]. Post performing this procedure, 2 of the disease classes, namely Asthma and LRTI, are found to have inadequate segmented samples for meaningful feature extraction and therefore, these two classes are not considered in our study. After these procedures, lung auscultation sounds from 87 out of 120 independent patients are usable. Table I represents data distribution at several levels of processing corresponding to the disease classes considered for this study.

| Disease Name | No. of unsegmented sound file | No. of Segmented and Filtered Sample | No. of Unique Patient | No. of Generalized Augmented Image |
|--------------|-------------------------------|-------------------------------------|-----------------------|-----------------------------------|
| Pneumonia    | 37                            | 41                                  | 3                     | 164                               |
| Bronchiectasis| 16                            | 55                                  | 6                     | 220                               |
| COPD         | 793                           | 13,63                               | 51                    | 2,016                             |
| Healthy      | 35                            | 42                                  | 13                    | 168                               |
| URTI         | 23                            | 21                                  | 8                     | 84                                |
| Bronchiolitis| 12                            | 65                                  | 6                     | 260                               |
| Total        | 917                           | 2,187                               | 87                    | 2,859                             |

C. Feature extraction

1) Empirical Mode Decomposition (EMD): EMD is a powerful self-adaptive signal decomposition method especially in the time scale and energy distribution aspects and highly suitable for analysis and processing of non-linear and non-stationary signals such as lung sounds and heart sounds [54]. It decomposes a given signal $x(t)$ into a finite set (N) of intrinsic mode functions, $IMF_1(t), IMF_2(t), \ldots, IMF_N(t)$, depending on the local characteristic time scale of the signal, with a view to expressing the original signal as the sum of all its IMF plus a final trend either monotonic or constant called residue, $r(t): x(t) = \sum_{i=1}^{N} IMF_i(t) + r(t)$ [55]. An IMF is a simple oscillatory function with the equal number of extrema and zero crossings and its envelopes must be symmetrical with respect to zero. Thus, the EMD detrends a signal and elicits underlying spectral patterns [54].

2) Continuous Wavelet Transform (CWT): Wavelet transform is defined as a signal processing method that can decompose a signal into an orthonormal wavelet basis or into a set of independent frequency channels [15], [29]. Using a basis function, i.e. the mother wavelet $g(t)$, and its scaled and
and analysis and segmentation of heart sounds, respectively, whereas our objective is to detect respiratory diseases from the lung auscultations.

E. Augmentation

The ICBHI 2017 dataset is highly imbalanced, with around 86% of the data belonging to COPD. Image augmentation using different color mapping schemes is employed to oversample the less represented classes and address the data imbalance issue [59]. Colormaps are three-column arrays containing RGB triplets where each row defines a distinct color. Scalogram representation using different color maps helps generalize the produced images.

From each of the audio samples of the less represented data classes, four scalograms are generated for each segmented sample using four different color mapping schemes: Parula, HSV, Jet, and Hot, which are available in MATLAB 2020a. A summary of segmented audio files and final augmented produced images.

Fig. 3. Box plot. (a) Scalogram using the conventional CWT approach; (b) Scalogram using the hybrid approach.

IV. PROPOSED LIGHTWEIGHT CNN ARCHITECTURE

CNN has become a popular approach for classifying image data, and recently there have been several works using CNN on classifying images produced from sounds [5], [31], [33]. However, due to memory constraints, a regular deep CNN model is computationally expensive with its large number of learnable parameters and arithmetic operations. Thus, it is not suitable for embedded devices as they cannot afford the processing complexity and storage space for parameters and weight values of filters [36]. Cloud computing methodology requires a higher RAM for this computationally intensive training and hence are outsourced [60]. For this reason, Lightweight CNN models are gaining popularity among researchers for their faster performance and compact size without compromising the much-needed accuracy performance compared to the well-known deep learning networks [61].

The architecture of the proposed CNN model consists of an input layer corresponding to the 3-channel input of 224x224 images. The architecture of the proposed model is illustrated in Fig. 4.
The 1st convolutional layer uses 64 output filters with a 55-pixel size kernel followed by a 22-pixel max-pooling layer. Three additional convolutional layers are stacked over the first layer, each having a 33-pixel size kernel with 64, 96, and 96 filters sequentially and corresponding batch-normalization and max-pooling layers with 22 pooling window. Outputs from all these layers are flattened and connected with five pairs of FC and dropout layers, followed by a SoftMax output layer with probability nodes for each class. ReLU activation layer is applied within convolution calculation and fully connected layers and employed to introduce nonlinearity within the calculation and reduce the time for convergence. It does not get activated for any negative value. Max pooling is used after the ReLU activation; it reduces the spatial dimensionality of the extracted feature maps, extracts the most important features, and is unaffected from locational bias [35]. To overcome the problem of more diverse data variance, the Batch Normalization layer with every convolution layer normalizes the extracted feature. It gives the network a representative power with a small number of parameters and faster training capability by reducing the variance.

V. Experimental Results

A. Evaluation Criteria

The augmented image sets are divided into 80% training and 20% validating parts for training and fine-tuning the model hyperparameters. Patient uniqueness, a critical aspect in the real-world applications, is maintained while dividing into training and validation parts as speaker dependency results in biased accuracy [46].

The classifier models’ performance is evaluated based on the well-known evaluation matrices, namely, accuracy, recall (sensitivity), precision, and F1-score. Additionally, specificity and ICBHI-score [38], [53], a dedicated metric involving both sensitivity and specificity to assess the performance of the frameworks using the ICBHI dataset, is used to evaluate the performance of our method.

B. Experimental Setup

The proposed CNN model is constructed using Keras and TensorFlow backend, and trained using NVidia K80 GPUs provided by Kaggle notebooks. The mini-batch training scheme is employed while feeding the image data into a model for tackling the class imbalance issue. This technique performs by oversampling the scarce classes while randomly undersampling a majority class. This strategy ensures that the CNN model takes an equal number of samples from each class during each of the training epochs and thereby forms a balanced training set [26].

The adaptive learning rate optimizer (Adam) with the learning rate of 0.00001 is used for compiling the model. The batch size needs to be a multiplier of 6 since an equal number of samples from each of the 3 and 6 data classes are taken in each training and validation batch [26]. In this study, batch size 6 has been taken for training and validation of both the classification schemes.

As stated earlier, both the ternary chronic classification (chronic, non-chronic, healthy) and six class (Bronchiectasis, Bronchiolitis, COPD, Healthy, Pneumonia, and URTI) pathological classification are carried out in this work. The classification performance of the proposed CNN model is compared with that of VGG16, a well-known CNN architecture for image classification [48] in both of the classification schemes. It should be noted that the experiments are performed using both the convention CWT-based scalogram and hybrid scalogram images. In addition, the performance of our proposed CNN model is compared with a number of well-known DL architectures such as VGG16 and AlexNet and several lightweight networks in terms of computational complexity and accuracy.

C. Classification Performance of the Proposed Framework

1) Chronic Classification: From Table II, it can be seen that using the hybrid scalogram method in conjunction with the proposed CNN model classifier shows the best accuracy, 99.21%. However, the corresponding accuracy obtained by using VGG16 is quite close (98.89%). Despite being a heavy model, the comparatively lower accuracy of VGG16 can be attributed to the over-fitting issue due to the limited number of images in different classes. When comparing conventional CWT scalogram to the proposed hybrid scalogram, considerable improvement in accuracy is evident for the latter using VGG16 and our proposed CNN model (9.5%-11.4%). The corresponding confusion matrices for both the models’ best results are illustrated in Fig. 5. The results depict that the
proposed method is better in ternary chronic classification than the VGG16.

2) Pathological Classification: For six-class Pathological classification, the proposed method involving hybrid scalogram and proposed CNN model classifier yields the best accuracy, 99.05%, as seen in Table II. Similar to the case in the ternary chronic classification scheme, the accuracy of VGG16 is slightly lower. However, since the dataset gets more segregated being divided into six different disease classes, the accuracy drop is more here. The proposed hybrid scalogram outperforms the conventional CWT scalogram with a larger margin (13.4%-14.7%) for both VGG16 and our proposed model. In general, the proposed method gives a better performance, which is apparent from the best confusion matrices shown in Fig. 6.

D. Comparison with Other Works

1) Respiratory Disease Classification: As discussed in section II, none of the existing works for respiratory disease classification explore the domain of patient-specific prediction. Some of the studies address the issues regarding class imbalance [39], [42]. Nevertheless, the extensive preprocessing, coupled with the ambiguous undersampling of the COPD disease class while oversampling all other disease classes, can complicate the reproducibility of [39]. Furthermore, in [42], FFT is applied to the entire respiratory sound signals, whereas our work focuses on segmented breath sounds. In our work, complete patient independence has been maintained in the train and validation set, which is not possible while using the entire lung auscultation signal due to the low number of samples. Therefore, our work aims to overcome all the drawbacks present in the existing methods. A comparison among the various methods, including the Proposed method, is provided in Table III.

It is observed that our proposed CNN model with the hybrid scalogram can perform on par with the existing state-of-the-art CNN and RNN models for both cases of classification while maintaining a patient independent train-validation scheme.

2) Computational Performance as a Lightweight Network: A detailed comparison is presented in Table IV among VGG16 [48], our proposed CNN model, AlexNet [49] and the existing state-of-the-art lightweight models such as MobileNetV2 [50], NASNet [51], ShuffleNetV2 [52] in terms of size, trainable parameters, the number of operations measured by multiply-add (MAdd) and accuracy on both chronic and pathological classification. It is observed that our proposed CNN model shows better results than VGG16 while requiring only 3% of the parameters. The proposed CNN model also outperforms the contemporary lightweight models, ShuffleNet V2, MobileNet V2, and NASNet relatively by 0.16%, 0.32%, and 0.80%, respectively, while obtaining better trade-off between the number of parameters, requiring significantly lower storage space and computational power. This makes our proposed lightweight model more suitable for real-time wearable
devices with faster and less resource-intensive training.

We have calculated the time required for the end-to-end classification of an auscultation sound using our framework. For this experiment, we only performed the preprocessing and inference step using all of our test data and calculated the mean and standard deviation of the required CPU time. We found that the preprocessing time for EMD+CWT is \(8s \pm 0.5s\), and only CWT is \(7.2s \pm 0.5s\). These processes are run on a Core i7 7500 processor with a 2.70-2.90GHz speed. Time required for the classification of a scalogram using the proposed network is \(0.07s \pm 0.01s\), while the MobileNetV2 takes \(0.085s \pm 0.01s\). Thus, the proposed CNN is faster in classifying a sound image as compared to MobileNetV2.

### TABLE IV

**Comparisons among Several Models from Lightweight Perspective**

| Parameter                  | VGG16  | AlexNet | MobileNet (v2) | ShuffleNet (v2) | NASNet |
|----------------------------|--------|---------|----------------|----------------|--------|
| Size (after training)      | 1.5GB  | 294MB   | 44.85MB        | 49MB           | 46.9MB |
| Trainable parameters       | 158M   | 25.704M | 3.7674M        | 4.2M           | 5.4M   |
| MAdd                      | 154.7G | 725M    | 371.93M        | 575M           | 564M   |
| Accuracy (6 classes)       | 97.60% | 98.23%  | 99.05%         | 98.73%         | 98.72% |
| Accuracy (3 classes)       | 97.60% | 98.51%  | 99.21%         | 98.72%         | 99.06% |

VI. CONCLUSION

In this work, we have proposed a lightweight CNN model to classify respiratory diseases using scalogram images of lung sounds. A hybrid approach employing both EMD and CWT is presented to generate the scalogram images. The publicly available ICBHI 2017 challenge dataset has been used for the Chronic and Pathological classification of respiratory diseases. The proposed method has provided a considerable accuracy of 99.21% for ternary chronic classification. In pathological classification among six disease classes, an accuracy of 99.05% is achieved. The obtained accuracies are higher than VGG16, which is a much larger network. In addition, for both cases of classifications, the proposed framework provides better or a comparable performance with respect to the existing state-of-the-art methods in terms of Precision, Recall, F1-score, Sensitivity, Specificity and ICBHI score. It is worthwhile to mention that unlike most of these methods, the classification performance of the proposed technique has been assessed, keeping the training and testing data-independent in terms of patients. The proposed classifier’s computational complexity has also been compared with a number of well-known CNN models and state-of-the-art lightweight networks. It has been shown to achieve high accuracy in classification while being a lightweight deep architecture. We believe that these attributes can enable the development of the automatic classification of respiratory diseases from lung auscultations in real-world clinical applications.
[61] B. Lim, B. Yang, and H. Kim, “Real-time lightweight cnn for detecting road object of various size,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, 2018, pp. 202–203.