IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 January 2024, accepted 31 January 2024, date of publication 5 February 2024, date of current version 13 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3361943

RESEARCH ARTICLE

Lung Sound Classification With Multi-Feature Integration Utilizing Lightweight CNN Model

THINIRA WANASINGHE[1], SAKUNI BANDARA[1], SUPUN MADUSANKA[1], DULANI MEEDENIYA[1], (Senior Member, IEEE), MEELAN BANDARA[1], AND ISABEL DE LA TORRE DÍEZ[2], (Member, IEEE)
[1]Department of Computer Science and Engineering, University of Moratuwa, Moratuwa 10400, Sri Lanka
[2]Department of Signal Theory and Communications, and Telematics Engineering, University of Valladolid, 47011 Valladolid, Spain

Corresponding author: Dulani Meedeniya (dulanim@cse.mrt.ac.lk)

This work was supported in part by the domain experts and in part by the University of Moratuwa, Sri Lanka, under the Senate Research Committee (SRC), Conference and Publishing Grant.

ABSTRACT Detecting respiratory diseases is of utmost importance, considering that respiratory ailments represent one of the most prevalent categories of diseases globally. The initial stage of lung disease detection involves auscultation conducted by specialists, relying significantly on their expertise. Therefore, automating the auscultation process for the detection of lung diseases can yield enhanced efficiency. Artificial intelligence (AI) has shown promise in improving the accuracy of lung sound classification by extracting features from lung sounds that are relevant to the classification task and learning the relationships between these features and the different pulmonary diseases. This paper utilizes two publicly available respiratory sound recordings namely, ICBHI 2017 challenge dataset and another lung sound dataset available at Mendeley Data. Foremost in this paper, we provide a detailed exposition about employing a Convolutional Neural Network (CNN) that utilizes feature extraction from Mel spectrograms, Mel frequency cepstral coefficients (MFCCs), and Chromagram. The highest accuracy achieved in the developed classification is 91.04% for 10 classes. Extending the contribution, this paper elaborates on the explanation of the classification model prediction by employing Explainable Artificial Intelligence (XAI). The novel contribution of this study is a CNN model that classifies lung sounds into 10 classes by combining audio-specific features to enhance the classification process.

INDEX TERMS Artificial intelligence, explainability, respiratory diseases, sound processing.

I. INTRODUCTION

Respiratory conditions are among the most prevalent medical ailments worldwide, affecting over 500 million individuals globally [1]. Most of the patients struggle to recognize or understand symptoms of chronic diseases, resulting in delayed diagnoses. The manual auscultation method is the most widely employed method by physicians to examine patients' lung sounds for disease diagnosis [2]. An expert physician is needed to auscultate the lung sound of the patient as it is complicated to diagnose diseases due to a lack of calibration of the instrument and also to the noisy environment like heartbeat sounds and coughing sounds [3]. Lung sound types are in 3 main categories; Wheeze, Crackles, and Rhonchi. Wheeze is a sustained, high-pitched, abnormal sound produced when there's a blockage in the airway, hindering normal breathing produced from the lungs of the patient with diseases such as pneumonia, and interstitial pulmonary fibrosis [4], [5]. Crackles are sudden, intermittent sounds that occur during both inhalation and exhalation associated with diseases like asthma and chronic obstructive pulmonary disease (COPD) [4], [5]. Rhonchi are continuous, low-pitched, and coarse sounds that resemble snoring or rattling produced due to the presence of airway secretions [5]. Utilizing these sound types through the auscultation process, medical practitioners gather their expertise on the relevant lung disease.

In recent years, a significant impact of AI in the healthcare sector, particularly in detecting conditions such as cancers, respiratory conditions, and neurological disorders

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

is evident [6], [7], [8], [9], [10], [11], [12]. Deep learning (DL) applications for lung sound classification have gained significant research interest [13], [14]. Deep learning feature extraction is a data-driven method that identifies features directly from raw data, facilitating the analysis of disease-specific features [15]. Convolutional Neural Networks (CNNs) have been used to classify spectrograms generated by lung sounds as they proficiently extract features from images, learning to recognize patterns, rendering them highly suitable for tasks like object detection, image segmentation, and classification [16], [17], [18].

Existing studies in lung sound classification have primarily focused on binary or small multi-class classification problems, often involving two to six classes [2], [19], [20], [21]. While these studies have contributed valuable insights into the applications of AI in healthcare, due to the constrained number of disease classes present in the datasets used in these studies, they may not fully address the diversity of diseases which is a challenge faced by clinicians. Employing explainable Artificial Intelligence (XAI) techniques in the prediction of the classification model is an under-explored area in the domain of lung sound classification. Compared to the existing studies, this study mainly addresses the challenge of classifying lung sounds into 10 classes as a novel contribution. Furthermore, we highlighted the improved model performance achieved by employing stacked audio inherent features compared to using individual features for feeding the CNN model. In addition, the integration of XAI techniques also adds a unique dimension to our approach.

The main contributions of the paper are as follows.

- Combining two publicly available datasets, namely the ICBHI 2017 challenge dataset [22] and the dataset developed by Fraiwan et al. [23], serves the purpose of increasing the variety of diseases covered in our study. Unlike many existing studies that rely on a single dataset, this approach enhances the generalizability of our model by incorporating a broader spectrum of respiratory conditions.
- Instead of using one type of audio feature, our paper suggests combining three different feature types: Mel Spectrogram, MFCC, and Chromagram for each audio sample. This approach creates a 3D representation of features, aiming to capture a richer set of characteristics from the lung sounds to improve the model performance.
- Develop a classification model using a CNN to classify 10 classes of lung-related diseases.
- To enhance the interpretability and trustworthiness of our classification model, we incorporate XAI techniques. This ensures transparency in the model's decision-making process, addressing concerns related to model trustworthiness.

The remainder of the paper is organized as follows; Section II describes the available public lung sound datasets, the taxonomy of techniques used in lung sound processing, and related works on lung sound classification for respiratory disease detection. In Section III, we delve into the comprehensive details of the developed model, the overall methodology, and XAI techniques. Section IV analyses the outcomes of the classification model and the results obtained through XAI. Section V discusses lessons learned during the research and state-of-the-art comparison in lung disease identification. Finally, Section VI presents the conclusion of the study.

II. BACKGROUND

A. LUNG SOUND DATASETS

Several studies have been reported employing DL techniques on automated respiratory sound classification to detect lung diseases, utilizing the existing publicly available datasets. While the majority of studies have focused on respiratory anomaly prediction, i.e., classifying lung sounds as wheeze, crackles, or rhonchi [5], [24], a lesser number of studies have stepped further on classifying lung sounds into various disease categories. However, the lack of a well-balanced audio data set of lung sounds has been a major challenge in automated respiratory sound classification.

ICBHI 2017 [22], which is a commonly used publicly available dataset, consists of 920 sound recordings. This dataset consists of 8 classes, where the duration ranges from 10s to 90s. Several studies have introduced new datasets with different augmentation techniques. For example, Tariq et al. [2] have applied an oversampling method that replicates samples to achieve balance among the classes of the ICBHI dataset. Additionally, the authors removed the asthma category, which contains only a single recording, from the dataset. Moreover, since the ICBHI dataset is relatively small for training DL models, Acharya and Basu [4] have employed augmentation techniques such as noise addition, speed variation, random shifting, and pitch shift, to increase the size of the dataset. In another point of view, as a contribution to the lack of publicly available datasets, Fraiwan et al. [23] have created a multiclass lung sound dataset of 112 entries, which contains audio data of 77 unhealthy subjects and 35 healthy subjects. Here, the duration of the records ranges from 5s to 30s and the dataset consists of 11 classes including healthy, pneumonia, asthma, COPD, lung fibrosis, heart failure, heart failure & lung fibrosis, heart failure & COPD, pleural effusion, asthma & lung fibrosis, and bronchitis.

Furthermore, there are other datasets created by research groups that are not publicly accessible [5], [25]. However, those recordings also consist of data imbalance issues, which is a common problem in the lung sound domain. The main reason for this is that certain diseases, such as common respiratory infections or conditions like asthma, are frequently encountered in clinical practice, whereas others, such as lung fibrosis and pleural effusion, are comparatively rare [26]. As a result, creating a lung sound dataset with an equal number of samples for each class becomes a challenging task.

B. TAXONOMY IN LUNG SOUND PROCESSING

The field of DL-based lung sound identification is evolving continuously. New approaches, techniques, and applications have been emerging over time. Figure 1 shows the taxonomy considered for this study, which is used to process, classify, analyze, and interpret the lung sound data using DL techniques. Based on the objectives of the application, researchers can select different combinations of techniques. Here, we considered data preprocessing approaches such as normalization [26], augmentation [3], [4], [27], feature extraction [3], [20], [27], [28], [29], classification [2], [20], [24], [30] and explainability [19], [31], [32], [33].

C. RELATED STUDIES

Literature has highlighted the role of preprocessing in sound classification, as it directly influences prediction accuracy [34]. Among different approaches, normalization plays an important role in data preprocessing to ensure consistency, especially when dealing with lung sounds recorded using various devices [34]. Ma et al. [26], have used the min-max normalization technique to process records in the ICBHI dataset to standardize the data across different recording devices.

From another point of view, the noise in the data, including heartbeat and coughing sounds can be utilized to simulate real-time scenarios in lung sound audio recordings [3], [27]. For example, Serbes et al. [28], have applied a 12th-order Butterworth band-pass filter with 120 and 1800 Hz cut-off frequencies, and Ma et al. [26], have used a 5th-order Butterworth band-pass filter, which helps to retain the frequency of interest from 100 to 2,000Hz to minimize noise effects.

Different augmentation techniques are used to maintain consistency in each class of the dataset [34]. Acharya et al. [4], have applied noise addition, speed variation, random shifting, and pitch shifting to create augmented samples to address the data imbalance and the lack of lung sound recordings data. The authors have stated that aside from increasing the dataset size, these data augmentation methods also help the network learn useful data representations despite different recording conditions, different equipment, patient age, and gender, and inter-patient variability of breathing rate. Similarly, Srivastava et al. [27], have immersed loudness augmentation, mask augmentation, shift augmentation, and speed augmentation to address the same issue. Additionally, the authors have trimmed and padded the audio files to a length of 20 seconds using the Python library Librosa [35].

Moreover, extracting features from audio data and feeding them into the classifiers have an impact on the accuracy of classifications. While the Mel-spectrogram and MFCC can be identified as the two most utilized spectrograms in related work, features such as Chromagram, Q-Chromagram, and Zero-crossing rate have also been employed to feed into the classifiers [27], [28]. For instance, Basu and Rana [20], have extracted MFCC (spectral features), from the audio data and 40 features have been extracted from each audio data to train the model, which has resulted in 95% accuracy. Similarly, Tariq et al. [2], have extracted three unique features from the audio samples, i.e., Spectrogram, MFCC, and Chromagram, to build the fusion of three optimal CNN models. In addition, several feature extraction and classification techniques for obstructive pulmonary diseases such as COPD and asthma are presented in the literature [29]. The process involves feature extraction through signals such as Fast Fourier Transform (FFT), Short Time Fourier Transform (STFT), spectrogram, and wavelet transform. As mentioned by Brunese et al. [3], they have used many spectrogram-based and non-spectrogram-based feature extraction techniques, including chromagram, root mean square, spectral centroid, MFCC, zero-crossing rate, spectral roll-off (SR), and other feature computations due to its demonstrated effectiveness in performing tasks involving supervised machine learning. Apart from the lung sound domain, other studies have also used techniques like Spectrogram, and wavelet transform [36].

In various disease categorizations, among diverse ML algorithms including SVM, kNN, and logistic regression, neural networks demonstrate notably superior prediction performance [3], [25], [37].

Accordingly, several research studies have explored the classification of lung sounds using DL algorithms. CNN is one of the most used and promising DL model for respiratory disease classification [27], [38], [39], [40]. Among them, Brunese et al. [3], have shown a possible approach to exploit a two-step classifier based on CNN to detect lung disease at a fine grain, to discriminate between healthy and affected lung conditions. Here, the abnormal sound detection by the first classifier is further categorized into 7 disease types using the second classifier with an F-score of 0.923. In another study, the lightweight CNN developed by Shuvo et al. [24], detected 5 types of lung diseases utilizing scalograms as the time-frequency representation of signals. Another CNN-based approach was developed by Basu et al. [20], by feeding MFCC features, to classify lung sounds into 5 disease categories and showed an accuracy and precision of 95.25% and 0.95, respectively. Tariq et al. [2], have built a fusion of three optimal CNN models by feeding the image feature vectors transformed from audio features to classify 6 disease types and have achieved an accuracy of 99%. In another study, Nguyen and Pernkopf [41], have introduced techniques like sample padding, feature splitting, a CNN snapshot ensemble, and a focal loss objective for lung sound classification, achieving superior performance with the highest ICBHI scores of 78.4% and 83.7% for the 4-class and 2-class tasks, respectively. Considering other classifiers, Perna et al. [30], have defined a learning framework based on Recurrent Neural Network (RNN) models to handle respiratory disease prediction problems at both anomaly and pathology levels to discover the time-dependent patterns from sound data. Moreover,

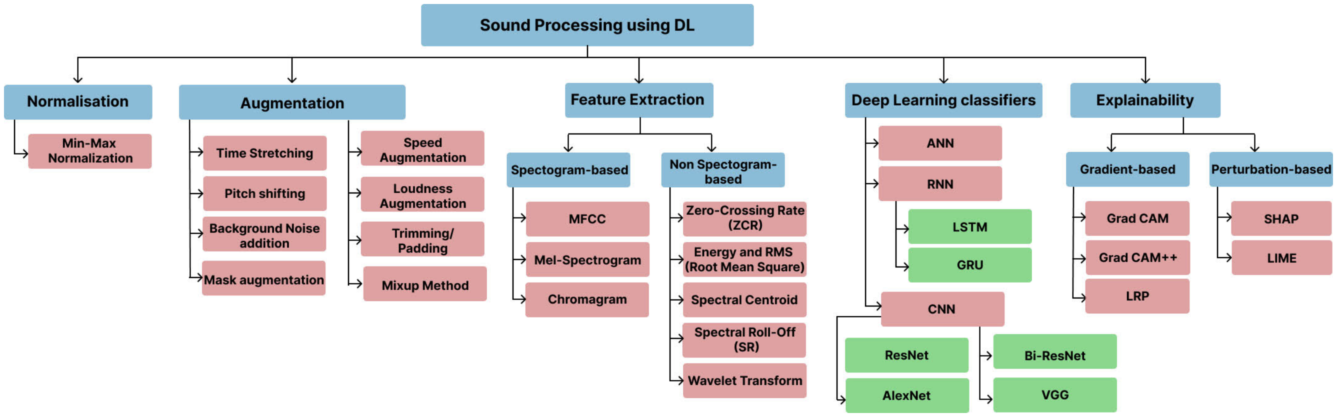


FIGURE 1. Taxonomy of Sound Preprocessing.

Nishi et al. [42] employed various classifiers, including support vector machine (SVM), k-nearest neighbour (KNN), logistic regression (LR), decision tree, and discriminant analysis (DA), for COPD identification. Notably, the SVM classifier yielded an initial accuracy of 83.6%. Subsequently, they successfully achieved a remarkable 100% accuracy. Although few studies have shown high accuracy levels [2], [20], some of them endure limitations such as lack of data and [4] and advanced feature extraction [28].

Furthermore, XAI techniques have been utilized in recent studies to improve the trustworthiness of the classifier and increase the confidence of the end-user in using DL-based support solutions [43]. Mainly, gradient-based methods [31], which consist of Gradient-weighted Class Activation Mapping (GradCAM), saliency maps, and perturbation-based methods [32], that include Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanation (SHAP) are used in the literature. However, only a few studies have addressed explainability in lung sound and heart sound classification domains. Additionally, few studies have expressed the value of interpretations of neural networks. Among them, Choi and Lee [19], have employed Grad-CAM to visualize the attention of a CNN model for lung disease diagnosis. The Grad-CAM visualizations showed that the model could focus on the characteristic points of the respiratory sounds for different diseases. Similarly, Topaloglu et al. [33], have used Grad-CAM to generate heat maps, effectively distinguishing asthma lung sounds from those of normal individuals. From another point of view, Wang et al. [44], have utilized SHAP values to explain the contribution of the different time-frequency representations for the output of the heart sound classification model based on the ImageNet CNN model.

Accordingly, in the domain of lung sound classification, the existing state-of-the-art research has primarily focused on the classification of up to seven distinct diseases. Furthermore, with respect to feature extraction from audio data, many studies have traditionally employed individual feature representations. This approach could limit the capacity of models, such as CNNs, to capture more spatial patterns within the data. The limited studies that have delved into explaining classification predictions have mainly relied on frequency-based interpretations. Therefore, considering the real-world practice, classifying lung sounds of a variety of conditions and interpreting the results in the original waveform would be of utmost importance.

III. SYSTEM MODEL

A. MATERIALS

In this study, we combined two publicly accessible datasets: ICBHI 2017 [22] and the dataset developed by Fraiwan et al. [23], both of which provide diagnoses for various lung diseases. We combined the two datasets to increase the number of available samples for specific lung conditions, and since each dataset has some different types of lung diseases, combining them gives us a bigger variety of classes to apply for the study.

The annotation files of ICBHI include seven disease types: chronic obstructive pulmonary disease (COPD), upper respiratory tract infection (URTI), asthma, lower respiratory tract infection (LRTI), bronchiectasis, pneumonia, bronchiolitis, and healthy lung sounds. However, the dataset exhibits a significant imbalance, where COPD accounts for approximately 86% of the data. The dataset created by Fraiwan et al. [23], consists of seven unique diseases, three classes representing combinations of these unique diseases, healthy samples, and, in total, it comprises 11 classes. As the primary step, we extracted and combined audio samples and standardized them into a fixed window length to ensure data consistency. The resulting dataset comprised 6-second audio clips, uniformly sampled at 44,100 Hz with 16-bit depth and a stereo channel configuration. The WavePad software, an open-source application for audio editing, was utilized to segment longer audio clips into multiple 6-second segments. This step became necessary for certain classes due to the insufficient availability of audio samples required to meet the predetermined range for each class.

We removed classes corresponding to "heart failure" from the dataset since they are unrelated to respiratory diseases and discarded the LRTI disease class to mitigate data imbalance. We further refined the dataset by excluding multi-labeled classes since they had less samples, resulting in the selection of 10 distinct classes: asthma, bronchiectasis, bronchiolitis, bronchitis, COPD, lung fibrosis, pleural effusion, pneumonia, URTI, and healthy. We augmented the two classes "pleural effusion" and "bronchitis" using the positive pitch shifting technique to balance the classes. The data volume before and after augmentation of two classes, subsequent to fixed window length audio file extraction from the combined dataset, is depicted in Table 1. Accordingly, we used a total of 1219 data records after augmentation, and using the 80:10:10 split ratio, the training, testing, and validation sets consist of 891, 111, and 112 records, respectively.

TABLE 1. Combined dataset: original and augmented data.

Category	Original data	Augmented data	Split of data with respect to 80:10:10 ratio		
			Training	Testing	Validation
Asthma	111	111	89	11	11
Bronchiectasis	105	105	84	11	11
Bronchiolitis	164	164	131	16	17
Bronchitis	53	106	85	10	11
COPD	115	115	92	11	12
Lung Fibrosis	114	114	91	11	12
Pleural Effusion	52	104	83	11	10
Pneumonia	160	160	128	16	16
URTI	101	101	81	10	10
Healthy	139	139	111	14	14

B. PROCESS VIEW

The primary goal of this research is to construct a model for the identification of pulmonary diseases from lung sound data. Figure 2 shows the overall process flow of this study. As the first step, we apply data augmentation techniques to mitigate the imbalance data issue and improve lung disease detection accuracy. Here, we applied a pitch shift of 1 semitone to each audio sample to minimize any impact on the original audio signal, using librosa.effects.pitch_shift library with a sample rate of 22050 and the number of steps as 1. Then normalization is applied to maintain numerical stability during processing and model training. The following step involves extracting three prominent audio features inherent from audio data: Mel-Spectrogram, MFCC, and Chromagram. Third, we stack the three feature types on top of each other to create a 3D feature representation for each audio sample. Fourth, we feed this 3D feature to our convolutional neural network. Finally, we employ Grad-CAM and Saliency to identify the relevant regions of the original waveform corresponding to the model's predictions.

C. FEATURE EXTRACTION

The feature extraction process is crucial in this study to transform the lung sound signals from the time-series domain to the time-frequency domain (spectrograms) and combine the extracted three time-frequency features to represent as images, which are fed to the image classification model. The spectrogram is calculated as in (1) and (2), where $S(\tau)$ is the time-domain signal, t is the time localization of short time Fourier transform ($STFT$), and $W(\tau - t)$ is a window function to cut and filter the signal [2], ω is the angular frequency and j is the imaginary unit, which is defined as the square root of -1.

$$spectrogram(t, \omega) = |STFT(t, \omega)|^2 \tag{1}$$

$$STFT(t, \omega) = \int_{-\infty}^{\infty} S(\tau).W(\tau - t).e^{-j\omega\tau} d\tau \tag{2}$$

We used techniques such as MFCC, Mel, and Chroma to generate the spectrograms for the input audio data. The Mel-Spectrogram represents the short-term power spectrum of a sound produced by sampling air pressure over time, transforming it from the time domain to the frequency domain using the Fast Fourier Transform (FFT), and then converting frequency to the Mel-scale and color dimension to amplitude [27]. The Mel scale is calculated as [2] and [45], where f is the frequency.

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

After computing the logarithm of the mel spectrogram values to compress the dynamic range, we get the log mel spectrogram. MFCCs are coefficients that collectively make up a Mel-frequency cepstrum (MFC) that represents the short-term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency [2]. The log mel spectrogram is subjected to a Discrete Cosine Transform (DCT) to decorrelate the coefficients and capture the most significant features. The resulting coefficients are the MFCCs. The n^{th} MFCC (C_n)) is given by (4), where M is the number of Mel filters, and log_mel(m) is the m^{th} value of the log mel spectrogram [46].

$$C_n = \sum_{m=0}^{M-1} cos\left(\frac{\pi n\,(2m+1)}{2M}\right) \cdot log\_mel\,(m) \tag{4}$$

The mel spectrogram is a more detailed representation of the power spectrum, capturing the distribution of energy across frequencies over time. In contrast, MFCCs are a more compact representation designed to emphasize characteristics relevant to human perception

Chroma features capture the harmonic and melodic characteristics of the sound [2]. Employing librosa.feature library, configuring n_fft and hop_length as 2048 and 512 respectively for each feature, we extracted the dominant features from each audio sample in the lung sound i.e., MFCC, Mel Spectrogram, and Chromargam with the dimension of (128×264). Following this, the created spectrograms for each

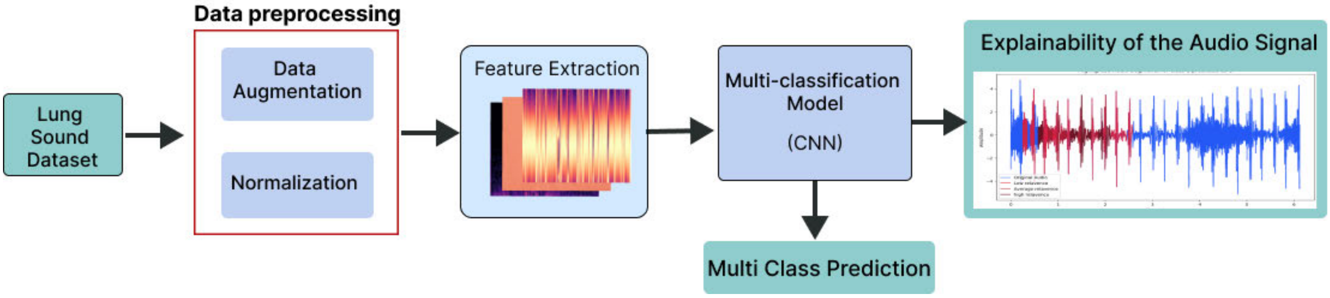


FIGURE 2. Overall Process.

audio sample were stacked using NumPy's 'stack' function, creating a three-dimensional input for the classification model (128×264×3). To encapsulate the 3D representation of features, each 2D feature (Mel, MfCC, Chroma), comparable to the RGB channels in an image, is stacked on top of each other to create three channels. This creates a 3D representation, with each channel capturing unique information from the audio signal. The resulting structure forms a comprehensive feature representation for each audio sample in the dataset. The overview of the feature engineering process is shown in figure 3, consisting of pre-processing and feature extraction stages, by taking an asthma sample as an example.

D. CLASSIFICATION MODEL

We proposed a customized CNN as the classifier of this study. CNNs are widely recognized and efficient DL models that learn and extract intricate features from image data [2], [34]. Our CNN model comprises a set of layers including the input, 2D convolutional, 2D max pooling, batch normalization, dropout, 2D global average pooling (GAP), and dense. The entry point of the model is the input layer, preserving that the model receives the input in a compatible format to process further. Convolutional 2D layers apply filters to identify spatial patterns, edges, and textures within the input, enabling the model to recognize relevant features through the ReLU activation function, as it enhances the model's ability to compute complex, non-linear relationships in the data [47]. The formula for the convolutional layer is expressed in ((5)) and ((6)). The 3-dimensional input tensor is represented by i, j, k while the output layer is represented by $y^{(l)}_{i,j,k}$. The weights for the filters are described by $w^{(l,f)}_{a,b,c}$ and σ is the activation function.

$$x^{(l)}_{i,j,k} = \sum_a \sum_b \sum_c w^{(l,f)}_{a,b,c} \cdot y^{(l-1)}_{i+a,j+b,k+c} + \text{bias}^{(f)} \quad (5)$$

$$y^{(l)}_{i,j,k} = \sigma\left(x^{(l)}_{i,j,k}\right) \quad (6)$$

The max pooling layer reduces the spatial dimension of the data by computing the maximum of the feature map, while the GAP layer calculates the average of feature maps across spatial dimensions, simplifying the spatial complexity of data to provide a compact representation for classification. Batch normalization enhances the training stability and speed of convergence by normalizing the input of each layer. The dropout layer implements a hold-out strategy to prevent overfitting by randomly deactivating a portion of neurons during training. The dense layer is pivotal for generating the model's final predictions, converting previously extracted features into class probabilities utilizing the 'softmax' activation function. Figure 4 illustrates the architecture of the model which is developed using Keras and a Tensorflow back-end. All the convolution layers consist of a kernel size of (3,3) with the same padding followed by a ReLU activation layer. Having the same padding allows input to be padded in such a way that the output feature map has the same spatial dimensions as the input. All the max pooling layers have a (2,2) window size.

To deploy on an embedded device, the CNN classification model must remain computationally efficient, avoiding excessive computational expense associated with a high number of learnable parameters and arithmetic operations [24]. To address this, the model has a lightweight architecture that has been optimized to substantially reduce the number of parameters to 510,825.

For model training, we assigned sample weights to balance the influence of different classes, ensuring that the model does not favour the majority classes and accurately predicts the minority classes. To compute sample weights for each class, we divided the number of recordings in the smallest class by the number of recordings in each class ensuring that the sample weight of the smallest class is set to one that has the highest importance. Moreover, the model was compiled using the Adam optimizer and adjusted the learning rate as 10^{-4}. The loss function employed for the multi-class classification task was Sparse Categorical Cross Entropy. It is an extension of the Cross-Entropy loss function that is used for binary classification problems [48]. The model computes the loss by comparing the predicted probability for each class with the corresponding actual class probability, penalizing the probability based on how far it is from the actual expected value. Cross-entropy is defined as in (7), where t_i is the truth label and p_i is the Softmax probability for the i^{th} class. The model was set to train for 100 epochs, with the integration of an early stopping callback function to prevent overfitting to the

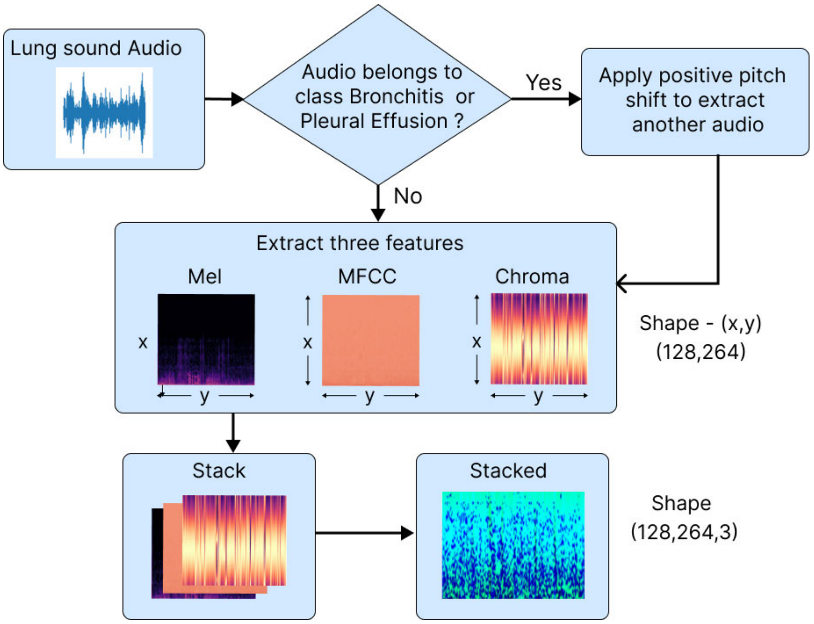


FIGURE 3. Feature engineering process (for an asthma sample).

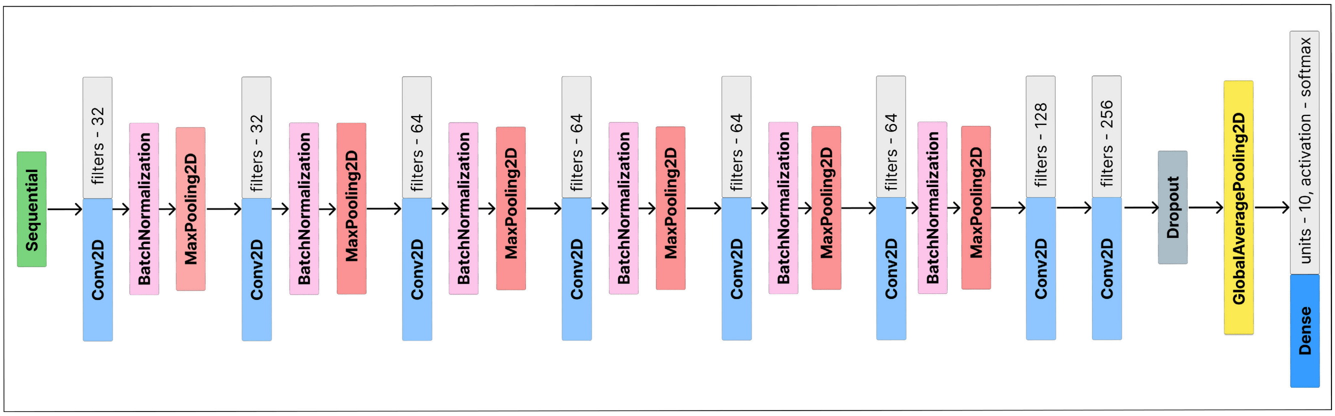


FIGURE 4. CNN architecture.

training data.

$$L_{CE} = -\sum_{i=1}^{n} t_i \, log(p_i), \;\; for\ n\ classes \qquad (7)$$

E. EXPLAINABILITY

The explainability of the proposed model is addressed using both Grad-CAM and the saliency method. The Grad-CAM, technique computes a weighted sum of the gradients of the last convolutional layer output with respect to the target class to visualize the input feature relevance to the prediction of the model [49]. Since the interpreted regions correspond to specific frequency bands, unlike pure image classifications, issues arise such as how the frequency band is related to the model's prediction. Next, to backtrack the interpretations of the model to analyze the behavior of the audio waveform, the saliency method was utilized. We employed two saliency methods, both using backpropagation to assess the relevance of individual features within the input layer.

The first method used the regular ReLU activation function, while the second used a guided ReLU function to mitigate the contribution loss of nodes resulting from the use of pooling layers [50]. Both methods provided values in the range of 0 to 1 representing the contribution levels of each input feature for a specific prediction. Next, we applied a threshold to mask out low-contributing features using the saliency map values and used these thresholds to identify the most contributive pixels in the image and mapped them back to the original waveform, highlighting the most relevant segments for the prediction. To further enhance the visualization, we calculated three distinct thresholds to indicate the contributions as low, average, and high, where the mean value of all the contributions derived from the saliency map is used to represent the average level while the ones above and below the average convey slow and high contributions, respectively. After calculating these threshold values, the contributed regions in the original audio are visualized using varying opacity levels to highlight areas of significance.

IV. RESULT ANALYSIS

We conducted comprehensive experiments for the proposed CNN model in four steps to assess that the stacked representation of the audio inherent features, i.e., Mel spectrogram, MFCC, and Chromagram is better in comparison to using these features individually for CNN classification. First, the model was trained and tested employing Mel Spectrograms. Next, the same step was repeated with MFCC and Chromagram. Finally, the experiment was done with the stacked representations. Furthermore, we experimented with the interpretation of our model's predictions using XAI techniques.

A. CLASSIFICATION RESULTS OF THE PROPOSED CNN

Our objective was to demonstrate that the combined feature representation outperforms the classification performance achieved using each of the features namely Mel-spectrogram, MFCC, and Chroma separately. Table 2 provides a summary of the results obtained for the proposed CNN model. We considered the weighted average as the testing accuracy for the models. The highest accuracy of 91.04% is obtained for the stacked features. The other performance criteria selected for evaluation were precision, recall, and F1-score as given in (8) to (11), along with the confusion matrix. Here, True Positive (TP) represent the outcome where the model correctly predicts the positive class. Similarly, a True Negative (TN) is an outcome where the model correctly predicts the negative class. Conversely, False Positive (FP) refer to cases where the model inaccurately predicts the positive class, and False Negative (FN) occur when the model incorrectly predicts the negative class.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

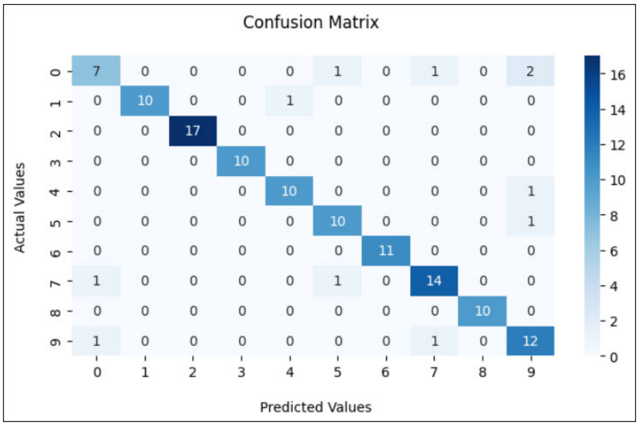
$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Reall} \tag{11}$$

TABLE 2. Results of classification.

Representation	Accuracy	F1-score	Precision	Recall
Mel Spectrogram	89.18%	0.883	0.883	0.885
MFCC	89.29%	0.891	0.901	0.893
Chromagram	68%	0.679	0.729	0.672
Stacked	91.04%	0.909	0.911	0.909

Since the dataset is imbalanced, we employed the weighted average for metric calculations since this approach accounts for variations in class frequencies. First, we assigned weights to each class based on their respective number of instances. Next, the weighted average was calculated by taking the sum of the product of the metric (accuracy, precision, recall, or F1-score) for each class and its corresponding weight, divided by the total sum of weights across all classes.

In order to enhance the analysis of the results, Figure 5 displays the confusion matrix for the approach with the highest reported weighted accuracy, which is the stacked representation. Here, the classes 0 to 9 denote the asthma, bronchiectasis, bronchiolitis, bronchitis, COPD, lung fibrosis, pleural effusion, pneumonia, URTI, and healthy classes, respectively. The confusion matrix serves as a visual representation and summary of the classification algorithm's performance. According to the matrix, the model has performed well for bronchiolitis, bronchitis, pleural effusion, and URTI classes and has an averagely high performance for bronchiectasis, COPD, and lung fibrosis, pneumonia, and healthy classes. The asthma class has obtained a relatively low accuracy with more false positives and false negatives. When considered overall, the model has shown impressive performance on the dataset, with significant precision and recall values.



Actual \ Predicted	0	1	2	3	4	5	6	7	8	9
0	7	0	0	0	0	1	0	1	0	2
1	0	10	0	0	1	0	0	0	0	0
2	0	0	17	0	0	0	0	0	0	0
3	0	0	0	10	0	0	0	0	0	0
4	0	0	0	0	10	0	0	0	0	1
5	0	0	0	0	0	10	0	0	0	1
6	0	0	0	0	0	0	11	0	0	0
7	1	0	0	0	0	1	0	14	0	0
8	0	0	0	0	0	0	0	0	10	0
9	1	0	0	0	0	0	0	1	0	12

FIGURE 5. Confusion Matrix for CNN Classification with Stacked Features, where classes denote 0 - asthma, 1 - bronchiectasis, 2 - bronchiolitis, 3 - bronchitis, 4 - COPD, 5 - lung fibrosis, 6 - pleural effusion, 7 - pneumonia, 8 - URTI, and 9- healthy class.

Figure 6 shows the training and loss curves on the CNN model. The overall decreasing trend in loss and increasing trend in accuracy depict that the CNN model is effectively learning and generalizing from the stacked features. Small spikes in both the loss and accuracy curves can be mostly due to the nature and variations present in the lung sound dataset.

In addition to the holdout method, we performed k-fold cross-validation with k=5 to further evaluate the robustness of our model. Across the five folds, our model demonstrated a strong performance, achieving an accuracy of 89.2%, an F1 score of 0.88, a precision of 0.89, and a recall of 0.88.

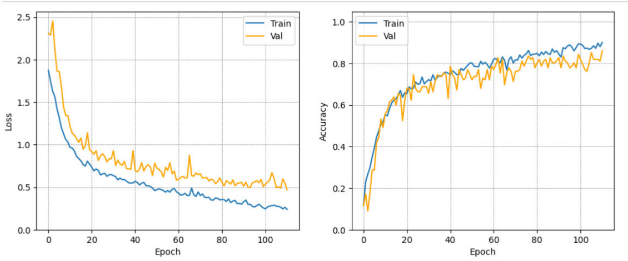


FIGURE 6. **Training and Loss curves.**

B. CLASSIFICATION RESULTS OF THE OTHER EXISTING CNN MODELS

To further emphasize on the performance of our model, we conducted an evaluation by comparing its performance with several existing CNN architectures, including Xception, DenseNet, MobileNetV2, InceptionV3, ResNet50, and VGG16. The results of this comparative analysis are presented in Table 3. Notably, when compared with the results in Table 3, our model demonstrates high performance and outperforms these models.

TABLE 3. **Results of classification.**

Model	Accuracy	F1-score	Precision	Recall
Xception	83.68%	0.831	0.843	0.836
DenseNet	88.65%	0.864	0.874	0.868
MobileNetV2	73.84%	0.722	0.739	0.721
InceptionV3	68.10%	0.672	0.689	0.680
ResNet50	79.33%	0.802	0.814	0.803
VGG16	80.58%	0.795	0.830	0.803

In order to analyze the results further, Figure 7 displays the confusion matrices for the above existing models.

While DenseNet has been able to perform well when considered class-wise, it has less accuracy in the healthy class. Xception, MobileNetV2, and VGG16 have an average accuracy in all the classes except the asthma class, while InceptionV3 and ResNet50 have poorly performed in lung fibrosis. Inception has the lowest performance when compared with all other models. So, the class-wise classification result is distributed across all the models, where an identical performance is not depicted. The confusion matrix of the proposed model as in Figure 5 shows better performance than DenseNet which outperformed other pre-trained models. Figure 6 shows the graphical comparison of all models using Receiver Operating Characteristic(ROC) curves. The ROC curve is a graphical representation used to assess the performance of a classification model and depicts the true positive rate against the false positive rate at each threshold setting. According to Figure 6, the Area Under the Curve of the proposed model is higher than the other existing models emphasizing the overall classification performance of the proposed model.

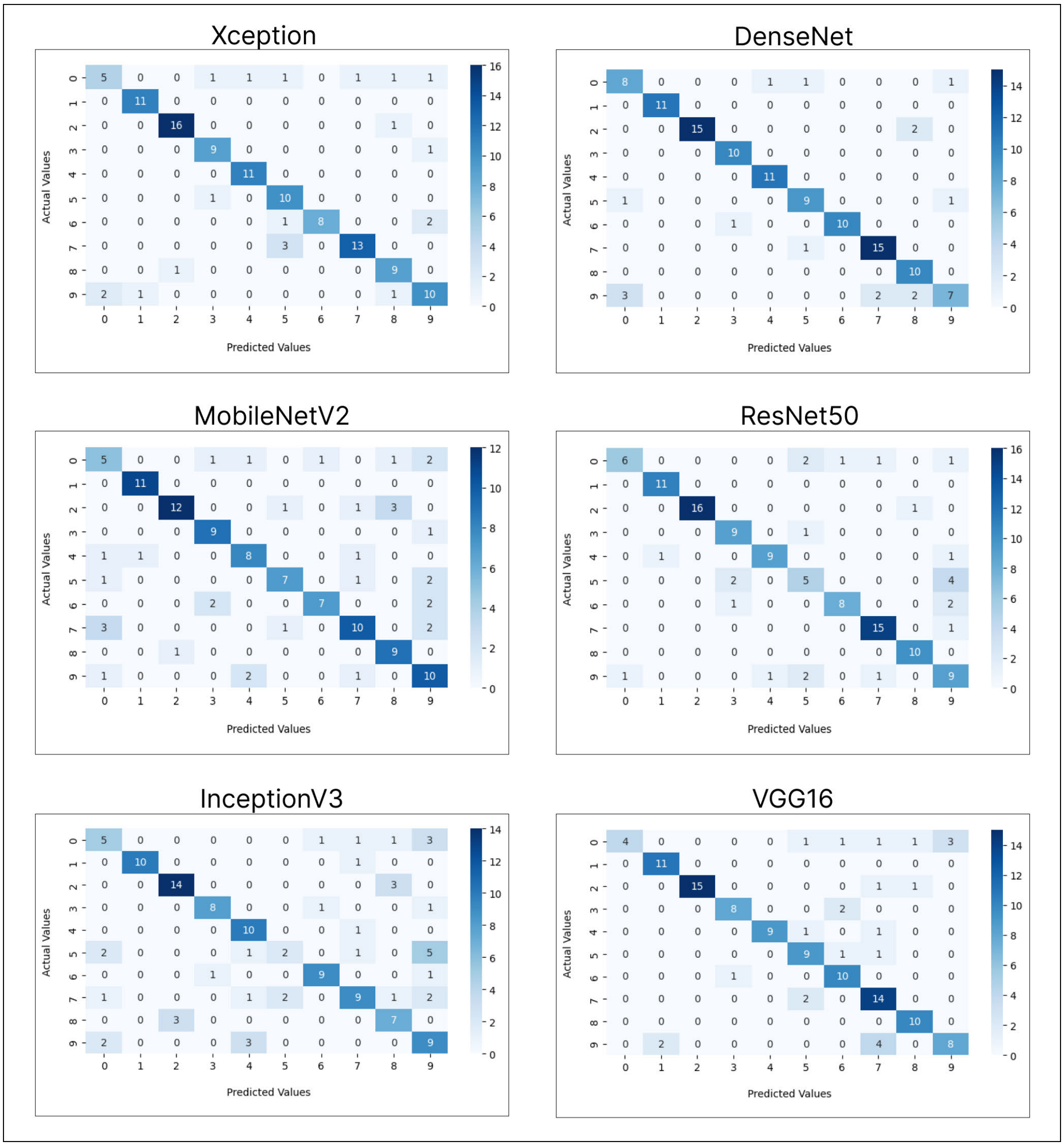
The architecture of the model developed by authors of [24], which has 3.7674M parameters and is claimed as lightweight, outperforms other contemporary lightweight models such as ShuffleNet V2, MobileNet V2, and NASNet, while obtaining better trade-off between the number of parameters, requiring significantly lower storage space and computational power. Our model, with 510,825 parameters, has a parameter count lower than that of [24], which needs much fewer computational requirements. We also calculated the inference time for a stacked image of a lung sound using a Core i7-8750 processor with clock speed specifications of 2.20 GHz. The time required for the classification of a stacked image using the proposed model is 0.026s ± 0.01s, while the MobileNetV2 takes 0.085s ± 0.01s and the lightweight model in [24] has taken 0.07s ± 0.01s. Thus, the proposed CNN has a lightweight architecture that is faster in classifying a sound image as compared to [24]. In addition, the model size of the proposed solution is 6.5 MB. Therefore, the solution gives a lightweight CNN model with multi-feature integration.

C. XAI RESULTS

In order to identify the best interpretable method for our study, we conducted experiments involving Grad-CAM heat maps and Saliency maps in two stages. At the initial stage, we generated the heat maps and the superimposed image of both the original feature representation and saliency maps as shown in Figure 9. Here, a clear relevance in features for model predictions could not be drawn since there was no mapping between the time-frequency characteristics of the audio and the explainable visualizations.

Accordingly, while both methods namely, Grad-CAM and saliency map showed color variances for relevant regions, the inability to identify any patterns within these images that would effectively support to identification of the disease was a clear challenge. As a solution for this, we sought a more direct means of identifying relevant areas using the original audio waveform. However, this requirement introduced another limitation with Grad-CAM heat maps that does not link the highlighted regions back to the original audio waveform. In order to address this limitation, further experiments were carried out with the saliency map approach. Here we made a significant breakthrough by highlighting important regions in the original waveform for better analysis with different relevance levels. When applied a guided ReLU activation function during the backpropagation process to assess the input relevance of stacked images, we observed improved interpretations than when utilizing regular ReLU activation.

Figure 10 depicts the XAI representation of bronchiolitis signal that is predicted as positive, utilizing regular ReLu and guided Relu. We have used the relevance values generated from the saliency map to get the most used pixels in the stacked spectrogram image, when representing the explainability of the prediction using the audio waveform. Since the x-axis of the spectrogram represents the time of the audio waveform, we utilized values in the time axis to highlight the audio waveform. Here, we used time regions



Xception

Actual \ Predicted	0	1	2	3	4	5	6	7	8	9
0	5	0	0	1	1	1	0	1	1	1
1	0	11	0	0	0	0	0	0	0	0
2	0	0	16	0	0	0	0	0	1	0
3	0	0	0	9	0	0	0	0	0	1
4	0	0	0	0	11	0	0	0	0	0
5	0	0	0	1	0	10	0	0	0	0
6	0	0	0	0	0	1	8	0	0	2
7	0	0	0	0	0	3	0	13	0	0
8	0	0	1	0	0	0	0	0	9	0
9	2	1	0	0	0	0	0	0	1	10

DenseNet

Actual \ Predicted	0	1	2	3	4	5	6	7	8	9
0	8	0	0	0	1	1	0	0	0	1
1	0	11	0	0	0	0	0	0	0	0
2	0	0	15	0	0	0	0	0	2	0
3	0	0	0	10	0	0	0	0	0	0
4	0	0	0	0	11	0	0	0	0	0
5	1	0	0	0	0	9	0	0	0	1
6	0	0	0	1	0	0	10	0	0	0
7	0	0	0	0	0	1	0	15	0	0
8	0	0	0	0	0	0	0	0	10	0
9	3	0	0	0	0	0	0	2	2	7

MobileNetV2

Actual \ Predicted	0	1	2	3	4	5	6	7	8	9
0	5	0	0	1	1	0	1	0	1	2
1	0	11	0	0	0	0	0	0	0	0
2	0	0	12	0	0	1	0	1	3	0
3	0	0	0	9	0	0	0	0	0	1
4	1	1	0	0	8	0	0	1	0	0
5	1	0	0	0	0	7	0	1	0	2
6	0	0	0	2	0	0	7	0	0	2
7	3	0	0	0	0	1	0	10	0	2
8	0	0	1	0	0	0	0	0	9	0
9	1	0	0	0	2	0	0	1	0	10

ResNet50

Actual \ Predicted	0	1	2	3	4	5	6	7	8	9
0	6	0	0	0	0	2	1	1	0	1
1	0	11	0	0	0	0	0	0	0	0
2	0	0	16	0	0	0	0	0	1	0
3	0	0	0	9	0	1	0	0	0	0
4	0	1	0	0	9	0	0	0	0	1
5	0	0	0	2	0	5	0	0	0	4
6	0	0	0	1	0	0	8	0	0	2
7	0	0	0	0	0	0	0	15	0	1
8	0	0	0	0	0	0	0	0	10	0
9	1	0	0	0	1	2	0	1	0	9

InceptionV3

Actual \ Predicted	0	1	2	3	4	5	6	7	8	9
0	5	0	0	0	0	0	1	1	1	3
1	0	10	0	0	0	0	0	1	0	0
2	0	0	14	0	0	0	0	0	3	0
3	0	0	0	8	0	0	1	0	0	1
4	0	0	0	0	10	0	0	1	0	0
5	2	0	0	0	1	2	0	1	0	5
6	0	0	0	1	0	0	9	0	0	1
7	1	0	0	0	1	2	0	9	1	2
8	0	0	3	0	0	0	0	0	7	0
9	2	0	0	0	3	0	0	0	0	9

VGG16

Actual \ Predicted	0	1	2	3	4	5	6	7	8	9
0	4	0	0	0	0	1	1	1	1	3
1	0	11	0	0	0	0	0	0	0	0
2	0	0	15	0	0	0	0	1	1	0
3	0	0	0	8	0	0	2	0	0	0
4	0	0	0	0	9	1	0	1	0	0
5	0	0	0	0	0	9	1	1	0	0
6	0	0	0	1	0	0	10	0	0	0
7	0	0	0	0	0	2	0	14	0	0
8	0	0	0	0	0	0	0	0	10	0
9	0	2	0	0	0	0	0	4	0	8

Actual Values
Predicted Values
Xception: 0 2 4 6 8 10 12 14 16
DenseNet: 0 2 4 6 8 10 12 14
MobileNetV2: 0 2 4 6 8 10 12
ResNet50: 0 2 4 6 8 10 12 14 16
InceptionV3: 0 2 4 6 8 10 12 14
VGG16: 0 2 4 6 8 10 12 14

FIGURE 7. Confusion matrices for Xception, DenseNet, MobileNetV2, InceptionV3, ResNet50, and VGG16, where classes denote 0 - asthma, 1 - bronchiectasis, 2 - bronchiolitis, 3 - bronchitis, 4 - COPD, 5 - lung fibrosis, 6 - pleural effusion, 7 - pneumonia, 8 - URTI, and 9- healthy class.

based on three threshold values high, average, and low to filter contribution into three levels in the audio waveform. These regions are indicated as blue for very low or zero contribution, yellow for low contribution, red for average contribution, and maroon for high contribution to the model's prediction. Here, different sections of the signal have contributed to the final predictions in different relevance levels, based on the extracted features and the computations performed in the layers of the proposed CNN classifier. It can be seen that, the guided-ReLu approach captured only the more important features for optimal prediction, compared to the regular-ReLu representation.

To evaluate our XAI approach, we employed important features from the saliency map as a mask, which we used to remove important features from the original spectrogram image. After masking important features from each stacked image in the test dataset to create a new masked dataset, this masked dataset is used to assess the original model. Table 4 shows the results of that evaluation.

TABLE 4. Explainability evaluation results.

Test Dataset	Accuracy	F1-score	Precision	Recall
Original test set	91.04%	0.909	0.911	0.909
Masked test set	70.49%	0.701	0.820	0.705

Based on the reduction of the accuracy, F1-score, precision, and recall with the masked test set when compared with values with the original test set, we can evaluate the explainability model. Since all the values reduce after the

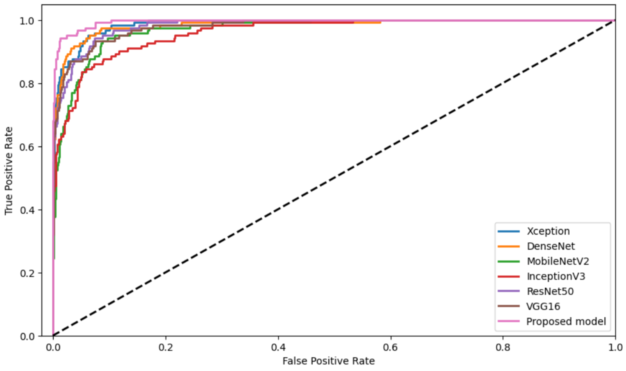


FIGURE 8. **ROC curves of existing models and proposed model.**

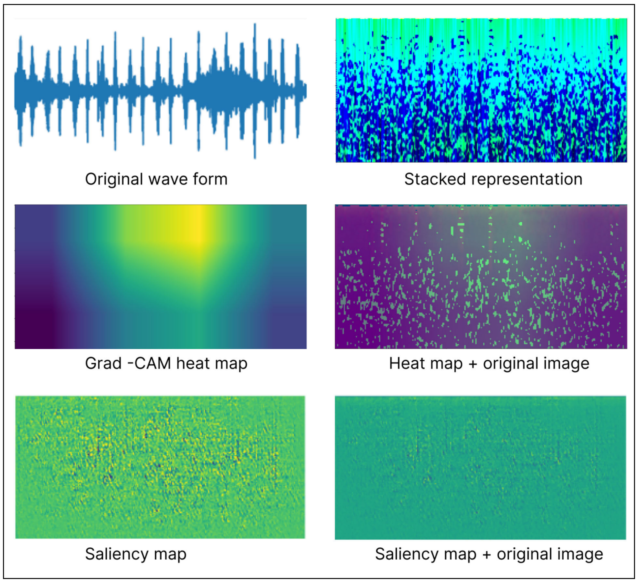


FIGURE 9. **Representation of heat maps and saliency maps for Bronchiectasis sample.**

masking it can be deduced that the explainability result gives the most important features used for prediction.

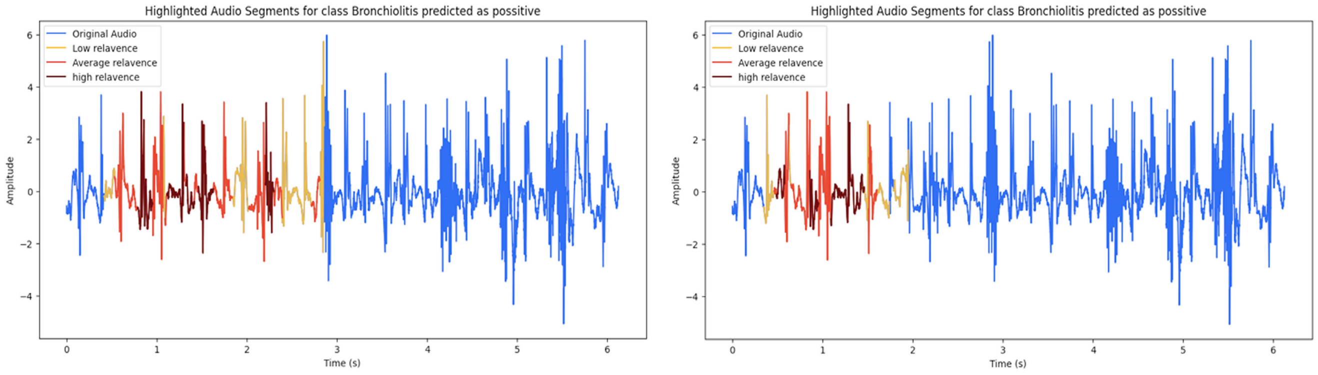
V. DISCUSSION

Throughout the research, we encountered various challenges and gained important insights into the domain of lung sound classification. As we evaluated the results of our experiments, we learned valuable lessons and were able to compare our work with the state-of-the-art research.

A. LESSONS LEARNED

In this study, we addressed lung sound multi-classification with 10 classes. We have shown the importance of stacking different feature extraction techniques to obtain better performances. The proposed CNN model classifies the lung sound data with an accuracy of 91.04%. Furthermore, we experienced different XAI techniques on sound data classification, which is another novel contribution in this domain. The observations of this study are described as follows.

- **Multi-class classification in lung sound domain**
  In the domain of lung sound classification, performing 10-class multi-class classification poses a unique set of challenges that significantly impact the accuracy and generalizability of the models developed. The main reason behind this is the limited availability of audio records. Also, it's often difficult to gather a dataset with an equal representation of various lung conditions, since some diseases are relatively rare while others are more frequent. Therefore, data imbalance will prevail almost all of the time. The highest accuracy showed by the proposed model, which is 91.04%, is up to the competing level while many of the existing studies yielded an accuracy of less than 90%.
- **Model validation by preventing data leakage**
  Model validation is the process of evaluating the performance and generalization ability of the DL model on new data. This ensures that the model can effectively make accurate predictions on unseen data that has not been seen during the training phase. Data leakage is another important aspect to be considered, which occurs when information from beyond the training data is utilized to train the model, enabling it to learn undisclosed details and thereby invalidating the estimated performance of the mode being constructed [51]. Preventing data leakage is crucial for reliable model validation, as it ensures the model is only evaluated based on what it learned during training, preventing any unintended exposure to external information that could compromise its ability to generalize accurately. According to the dataset details shown in Table 1, we split the dataset into training, testing, and validation sets in such a way that no subject contributed data to both the training and validation sets. Thus, our proposed solution has performed model validation avoiding any data leakages, and ensuring that there are no overoptimistic results.
- **Importance of feature representation methodology**
  When utilizing a CNN for audio classification, extracting audio inherent features that best represent the audio signal has extreme importance. Our approach showed that combining Mel Spectrogram, MFCC, and Chromagram by stacking on top of each other yields better results. According to our observations, the main reason behind this is the ability of the CNN model to explore as many feature patterns as possible through its convolutional layers and adjust weights for better classification. When considering features individually for classification, it can be seen that Chromagram yields very low results while Mel Spectrogram and MFCC perform almost equally.
- **Importance of retaining noise in lung sounds**
  In the context of lung sound classification, removing noise is not always an appropriate preprocessing step. Since the lung sound databases utilized by this study



Regular ReLU

Guided ReLU

FIGURE 10. Comparison between regular ReLU and Guided ReLU.

were made by collecting records of respiratory sounds from different research teams with different recording tools, the recordings can often exhibit various types of noises, mirroring real-world scenarios [52]. Some studies have considered adding random noise to the audio samples to avoid overfitting during training [2]. Since a certain level of uniformity in the noise characteristics cannot be determined in the datasets, conducting experiments in different signal-to-noise ratio levels was not considered.

- **Importance of utilizing proper XAI techniques**
The explainability results showed two key observations. First, unlike pure image classification tasks, interpretability techniques applied to audio data, such as spectrograms, often fall short of providing effective insights for human analysis. For audio, understanding the impact of frequency content alone is not sufficient. Instead, we found that backtracking and highlighting specific regions in the audio waveform offer a more insightful approach since it enables analysis of waveform shapes and the time intervals within the highlighted regions. Many image classification XAI evaluation techniques, including image entropy and pixel flipping, predominantly focus on assessing explainability in the context of human-readable images. Since our approach involves classifying audio waves using spectrogram images, these conventional evaluation techniques are ineffective. However, it is important to note that even these methods require further improvement since we are currently adapting XAI methods primarily designed for image processing to the audio domain.

B. COMPARISON WITH STATE-OF-THE-ART RESEARCH

We conducted a comprehensive evaluation of the lung sound classification models, regarding feature selection and representation, model architecture, dataset used, number of classes, performance results, and XAI techniques utilized. The overall comparison of our study for lung sound classification with state-of-the-art research is shown in Table 5. The best state-of-the-art approach for lung classification that has obtained the highest accuracy of 99% is by [2]. However, they have augmented the data about ten times the original and categorized it into 6 classes. The study by Chen et al. [53], has used different features using the ResNet-50, with over 23M trainable parameters resulting in 98.79% accuracy, but their classification only includes 3 classes. Our model has shallow trainable parameters, thus it enables us to run the model in resource-constrained environments without sacrificing accuracy. Choi and Lee [19], have used CNN with attention module and depthwise separable convolution to classify 6 classes with 92.5% accuracy and employed Grad-CAM as an XAI technique to visualize the important regions of the input data that contribute to the model's decision. Although the accuracy result published in this paper does not outperform the other state-of-the-art performances, a direct comparison cannot be drawn since we perform a 10-class classification and employ several XAI methods.

TABLE 5. Comparison of existing studies.

Study	Model	Features	Classes #	Result	XAI
[2]	CNN	Mel/ MFCC/ Chroma	6	99%	None
[53]	ResNet50	Spectro.	3	98.79%	None
[19]	CNN + attention module	Mel	6	92.56%	Grad-CAM
[20]	CNN	MFCC	5	95.67%	None
[21]	CNN	Spectro.	4	71.15%	None
Our study	**CNN**	**(Stacked) Mel/ MFCC/ Chroma.**	**10**	**91.04%**	**Saliency, Grad-CAM**

C. CHALLENGES AND FUTURE EXTENSIONS

Several research directions could be explored as future research in the lung sound domain. Developing quantized models employing techniques such as quantization-aware training, and post-quantization techniques will be of an importance when proceeding with the research work for real-time automatic lung sound auscultations. The generalizability of the model should be tested by deploying the system in an edge device, without sacrificing accuracy which is not a complex task considering the low number of parameters. However, proper quantization techniques should be employed for this task. An improvement is needed in pre-processing or data augmentation techniques to overcome data issues such as data imbalance and noise, which are common in medical research, resulting in poor performance. Moreover, exploring interpretable methods for audio classifications needs proper attention from researchers towards the improvement of this domain. Also, addressing data issues in medical research will remain a prominent focus of research work for an extended period. Furthermore, a validated model using medical practitioners can be implemented as a support tool for clinical settings. A mobile application comprising the developed model can be developed, which is connected to an electronic stethoscope to listen to lung sounds and predict lung diseases in real-time. Thus, the generalizability of this model can be improved by conducting an external validation process with new patients' data in clinical practice [54].

VI. CONCLUSION

This study presents a deep learning approach for lung sound classification. As a novel contribution, we proposed a CNN model to classify lung sounds into 10 classes employing stacked features as a means to improve the performance of the model. We utilized two publicly available lung sound datasets with different numbers of samples and class imbalance ratios. By employing stacked feature representation, our CNN model achieved a maximum accuracy of 91.04%, hence showing the importance of combining different audio inherent features to discover new patterns of features involved in a specific disease in the lung sound domain. Explainability methods such as Saliency maps and Grad-CAM provided insights into the classification model's predictions allowing us to visualize the important regions in the audio waveform, the time intervals, and frequency content. We will further improve the research by exploring more DL techniques such as Quantized CNNs, and Quantum CNNs to improve the model performance and implement a real-time lung sound analysis system.

REFERENCES

[1] J. B. Soriano, P. J. Kendrick, K. R. Paulson, V. Gupta, E. M. Abrams, R. A. Adedoyin, T. B. Adhikari, S. M. Advani, A. Agrawal, and E. Ahmadian, "Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: A systematic analysis for the global burden of disease study 2017," *Lancet Respiratory Med.*, vol. 8, no. 6, pp. 585–596, 2020.

[2] Z. Tariq, S. K. Shah, and Y. Lee, "Feature-based fusion using CNN for lung and heart sound classification," *Sensors*, vol. 22, no. 4, p. 1521, Feb. 2022.

[3] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "A neural network-based method for respiratory sound analysis and lung disease detection," *Appl. Sci.*, vol. 12, no. 8, p. 3877, Apr. 2022.

[4] 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.

[5] Y. Kim, Y. Hyon, S. S. Jung, S. Lee, G. Yoo, C. Chung, and T. Ha, "Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning," *Sci. Rep.*, vol. 11, no. 1, p. 17186, Aug. 2021.

[6] N. Wijethilake, D. Meedeniya, C. Chitraranjan, I. Perera, M. Islam, and H. Ren, "Glioma survival analysis empowered with data engineering—A survey," *IEEE Access*, vol. 9, pp. 43168–43191, 2021.

[7] M. T. Nguyen, W. W. Lin, and J. H. Huang, "Heart sound classification using deep learning techniques based on log-mel spectrogram," *Circuits, Syst., Signal Process.*, vol. 42, no. 1, pp. 344–360, Jan. 2023.

[8] D. Meedeniya, H. Kumarasinghe, S. Kolonne, C. Fernando, I. D. L. T. Díez, and G. Marques, "Chest X-ray analysis empowered with deep learning: A systematic review," *Appl. Soft Comput.*, vol. 126, Sep. 2022, Art. no. 109319.

[9] M. Xiang, J. Zang, J. Wang, H. Wang, C. Zhou, R. Bi, Z. Zhang, and C. Xue, "Research of heart sound classification using two-dimensional features," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104190.

[10] L. Gamage, U. Isuranga, S. De Silva, and D. Meedeniya, "Melanoma skin cancer classification with explainability," in *Proc. 3rd Int. Conf. Adv. Res. Comput. (ICARC)*, Belihuloya, Sri Lanka, Feb. 2023, pp. 30–35.

[11] D. Meedeniya and I. Rubasinghe, "A review of supportive computational approaches for neurological disorder identification," in *Interdisciplinary Approaches to Altering Neurodevelopmental Disorders*. Hershey, PA, USA: IGI Global, 2020, pp. 271–302.

[12] N. Faruqui, M. A. Yousuf, M. Whaiduzzaman, A. K. M. Azad, A. Barros, and M. A. Moni, "LungNet: A hybrid deep-CNN model for lung cancer diagnosis using CT and wearable sensor-based medical IoT data," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 104961.

[13] L. Fraiwan, O. Hassanin, M. Fraiwan, B. Khassawneh, A. M. Ibnian, and M. Alkhodari, "Automatic identification of respiratory diseases from stethoscopic lung sound signals using ensemble classifiers," *Biocybern. Biomed. Eng.*, vol. 41, no. 1, pp. 1–14, Jan. 2021.

[14] Z. Zhao, Z. Gong, M. Niu, J. Ma, H. Wang, Z. Zhang, and Y. Li, "Automatic respiratory sound classification via multi-branch temporal convolutional network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 9102–9106.

[15] D.-M. Huang, J. Huang, K. Qiao, N.-S. Zhong, H.-Z. Lu, and W.-J. Wang, "Deep learning-based lung sound analysis for intelligent stethoscope," *Mil. Med. Res.*, vol. 10, no. 1, p. 44, Sep. 2023.

[16] M. Tripathi, *Image Processing Using CNN: Beginner's Guide to Image Processing*. Gurgaon, India: Analytics Vidhya, 2021.

[17] C. A. Dimoulas, "Audiovisual spatial-audio analysis by means of sound localization and imaging: A multimedia healthcare framework in abdominal sound mapping," *IEEE Trans. Multimedia*, vol. 18, no. 10, pp. 1969–1976, Oct. 2016.

[18] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *Int. J. Autom. Comput.*, vol. 18, no. 3, pp. 351–376, Jun. 2021.

[19] Y. Choi and H. Lee, "Interpretation of lung disease classification with light attention connected module," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104695.

[20] 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.

[21] F. Demir, A. M. Ismael, and A. Sengur, "Classification of lung sounds with CNN model using parallel pooling structure," *IEEE Access*, vol. 8, pp. 105376–105383, 2020.

[22] (Jul. 10, 2017). *ICBHI 2017 Challenge*. [Online]. Available: https://bhichallenge.med.auth.gr/ICBHI_2017_Challenge

[23] 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.

[24] 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.

[25] M. A. Islam, I. Bandyopadhyaya, P. Bhattacharyya, and G. Saha, "Multichannel lung sound analysis for asthma detection," *Comput. Methods Programs Biomed.*, vol. 159, pp. 111–123, Jun. 2018.

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

[27] A. Srivastava, S. Jain, R. Miranda, S. Patil, S. Pandya, and K. Kotecha, "Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease," *PeerJ Comput. Sci.*, vol. 7, p. e369, Feb. 2021.

[28] G. Serbes, S. Ulukaya, and Y. P. Kahya, "An automated lung sound preprocessing and classification system based onspectral analysis methods," in *Proc. Int. Conf. Biomed. Health Inform.*, Thessaloniki, Greece. Cham, Switzerland: Springer, Nov. 2017, pp. 45–49.

[29] R. Dubey and R. M. Bodade, "A review of classification techniques based on neural networks for pulmonary obstructive diseases," *Proceedings of Recent Advances in Interdisciplinary Trends in Engineering & Applications (RAITEA)*. Indore, India: IPS Academy, 2019.

[30] D. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *Proc. IEEE 32nd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2019, pp. 50–55.

[31] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," 2017, *arXiv:1711.06104*.

[32] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, pp. 68–77, Dec. 2019.

[33] I. Topaloglu, P. D. Barua, A. M. Yildiz, T. Keles, S. Dogan, M. Baygin, H. F. Gul, T. Tuncer, R.-S. Tan, and U. R. Acharya, "Explainable attention ResNet18-based model for asthma detection using stethoscope lung sounds," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 106887.

[34] D. Meedeniya, *Deep Learning: A Beginners' Guide*. Boca Raton, FL, USA: CRC Press, 2023.

[35] (Sep. 24, 2023). *Feature Extraction Librosa 0.10.1 Documentation*. [Online]. Available: https://librosa.org/doc/latest/feature.html

[36] D. Kaplun, A. Voznesensky, S. Romanov, V. Andreev, and D. Butusov, "Classification of hydroacoustic signals based on harmonic wavelets and a deep learning artificial intelligence system," *Appl. Sci.*, vol. 10, no. 9, p. 3097, Apr. 2020.

[37] A. Monaco, N. Amoroso, L. Bellantuono, E. Pantaleo, S. Tangaro, and R. Bellotti, "Multi-time-scale features for accurate respiratory sound classification," *Appl. Sci.*, vol. 10, no. 23, p. 8606, Dec. 2020.

[38] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F.-M. Smolle-Jüttner, H. Olschewski, and F. Pernkopf, "Multi-channel lung sound classification with convolutional recurrent neural networks," *Comput. Biol. Med.*, vol. 122, Jul. 2020, Art. no. 103831.

[39] K. K. Lella and A. Pja, "Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: Cough, voice, and breath," *Alexandria Eng. J.*, vol. 61, no. 2, pp. 1319–1334, Feb. 2022.

[40] C. Fernando, S. Kolonne, H. Kumarasinghe, and D. Meedeniya, "Chest radiographs classification using multi-model deep learning: A comparative study," in *Proc. 2nd Int. Conf. Adv. Res. Comput. (ICARC)*, Belihuloya, Sri Lanka, Feb. 2022, pp. 165–170.

[41] T. Nguyen and F. Pernkopf, "Lung sound classification using snapshot ensemble of convolutional neural networks," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Montreal, QC, Canada, Jul. 2020, pp. 760–763.

[42] N. S. Haider, B. K. Singh, R. Periyasamy, and A. K. Behera, "Respiratory sound based classification of chronic obstructive pulmonary disease: A risk stratification approach in machine learning paradigm," *J. Med. Syst.*, vol. 43, no. 8, pp. 1–13, Aug. 2019.

[43] S. Wickramanayake, S. Rasnayaka, M. Gamage, D. Meedeniya, and I. Perera, *Explainable Artificial Intelligence for Enhanced Living Environments: A Study on User Perspective* (Advances in Computers). Amsterdam, The Netherlands: Elsevier, 2023.

[44] Z. Wang, K. Qian, H. Liu, B. Hu, B. W. Schuller, and Y. Yamamoto, "Exploring interpretable representations for heart sound abnormality detection," *Biomed. Signal Process. Control*, vol. 82, Apr. 2023, Art. no. 104569.

[45] M. Tanveer and R. B. Pachori, *Machine Intelligence and Signal Analysis*, vol. 748. Cham, Switzerland: Springer, 2019.

[46] *Discrete Cosine Transform MATLAB & Simulink*. Accessed: Dec. 20, 2023. [Online]. Available: https://www.mathworks.com/help/images/discrete-cosine-transform.html

[47] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, and H. Ghayvat, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021.

[48] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.

[49] 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.

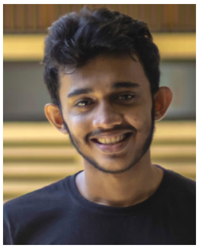
[50] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[51] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Sep. 2023, Art. no. 100804.

[52] G. Chambres, P. Hanna, and M. Desainte-Catherine, "Automatic detection of patient with respiratory diseases using lung sound analysis," in *Proc. Int. Conf. Content-Based Multimedia Indexing (CBMI)*, La Rochelle, France, Sep. 2018, pp. 1–6.

[53] H. Chen, X. Yuan, Z. Pei, M. Li, and J. Li, "Triple-classification of respiratory sounds using optimized S-Transform and deep residual networks," *IEEE Access*, vol. 7, pp. 32845–32852, 2019.

[54] R. D. Riley, L. Archer, K. I. E. Snell, J. Ensor, P. Dhiman, G. P. Martin, L. J. Bonnett, and G. S. Collins, "Evaluation of clinical prediction models (Part 2): How to undertake an external validation study," *Brit. Med. J.*, vol. 384, Jan. 2024, Art. no. e074820.



THINIRA WANASINGHE received the B.Sc. degree in engineering from the Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka. His main research interests include deep learning and computer vision.

SAKUNI BANDARA received the B.Sc. degree in engineering from the Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka. Her main research interests include artificial intelligence, data science, and deep learning.

SUPUN MADUSANKA received the B.Sc. degree in computer science and engineering from the University of Moratuwa, Sri Lanka, with a focus on frontend development, UI design, and a strong interest in machine learning and explainable AI.

DULANI MEEDENIYA (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of St Andrews, U.K. She is currently a Professor of computer science and engineering with the University of Moratuwa, Sri Lanka. She is also the Director of the Bio-Health Informatics Group, where she engages in many collaborative research. She is a coauthor of more than 100 publications in indexed journals, peer-reviewed conferences, and international book chapters. Her main research interests include software modeling and design, bio-health informatics, deep learning, and technology-enhanced learning. She is a fellow of HEA, U.K., MIET, a member of ACM, and a Chartered Engineer registered at EC, U.K. She serves as a reviewer, a program committee, and an editorial team member for many international conferences and journals.

MEELAN BANDARA received the B.Sc. degree in engineering from the Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka. His main research interests include environmental sound classification, deep learning-based audio classification, blockchain infrastructure, and capital market infrastructure.

ISABEL DE LA TORRE DÍEZ (Member, IEEE) is currently a Professor with the Department of Signal Theory and Communications, University of Valladolid, Spain, where she is the Leader of the GTe Research Group (http://sigte.tel.uva.es). She is the author or coauthor of more than 210 papers in SCI journals, peer-reviewed conference proceedings, books, and international book chapters. She has coauthored 21 registered innovative software. She has been involved in more than 100 program committees of international conferences, until 2021. She has participated/coordinated in 44 funded European, national, and regional research projects. Her research interests include e-health, telemedicine, m-health, sensors, data mining, cloud, quality of service (QoS), quality of experience (QoE), and economical evaluation of e-health services and apps.