Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Biomedical Signal Processing and Control 84 (2023) 104695

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Interpretation of lung disease classification with light attention connected module

Youngjin Choi, Hongchul Lee *

School of Industrial Management Engineering, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea

ARTICLE INFO

Keywords:
Respiratory sound
Lung disease
Attention
ECA-Net
eXplainable AI
Grad-CAM

ABSTRACT

Lung diseases lead to complications from obstructive diseases, and the COVID-19 pandemic has increased lung disease-related deaths. Medical practitioners use stethoscopes to diagnose lung disease. However, an artificial intelligence model capable of objective judgment is required since the experience and diagnosis of respiratory sounds differ. Therefore, in this study, we propose a lung disease classification model that uses an attention module and deep learning. Respiratory sounds were extracted using log-Mel spectrogram MFCC. Normal and five types of adventitious sounds were effectively classified by improving VGGish and adding a light attention connected module to which the efficient channel attention module (ECA-Net) was applied. The performance of the model was evaluated for accuracy, precision, sensitivity, specificity, f1-score, and balanced accuracy, which were 92.56%, 92.81%, 92.22%, 98.50%, 92.29%, and 95.4%, respectively. We confirmed high performance according to the attention effect. The classification causes of lung diseases were analyzed using gradient-weighted class activation mapping (Grad-CAM), and the performances of their models were compared using open lung sounds measured using a Littmann 3200 stethoscope. The experts' opinions were also included. Our results will contribute to the early diagnosis and interpretation of diseases in patients with lung disease by utilizing algorithms in smart medical stethoscopes.

1. Introduction

Chronic obstructive pulmonary disease (COPD) ranked third worldwide in 2019, caused more than 3.2 million deaths, according to World Health Statistics [1]. Furthermore, more than one billion people worldwide suffer from COPD, asthma, occupational lung disease, and acute lower respiratory tract infections [2]. Although obstructive pulmonary disease is simple to diagnose, it causes pneumonia complications due to COVID-19, and the infectious disease is further spreading owing to the lack of professional medical personnel. Particularly, when a COPD patient becomes infected with COVID-19, the outcome becomes worse than general patients. Many patients with underlying COPD were hospitalized in the intensive care unit (ICU), and 65 % of severely ill patients died with a history of COPD [3]. Additionally, lung diseases, such as asthma and pneumonia, have been the leading causes of death since COVID-19, and considerable research and efforts are being invested toward the early diagnosis of lung diseases [4,5]. Therefore, the early detection of diseases is important to prevent epidemics and lung diseases. Early detection of respiratory diseases can reduce the chances of complications with proper diagnosis and treatment [6].

Medical staff use X-rays, computed tomography (CT), and stethoscopes to diagnose lung diseases [2]. In particular, stethoscopes are widely used for lung auscultation because it is possible to observe the characteristic sound of breath [7]. Auscultation is noninvasive and inexpensive, making it easy to observe various lung diseases. Respiratory sounds measured by auscultation were divided into normal and adventitious respiratory sounds. Since adventitious sounds are distinguished from normal sounds, they are useful for diagnosing diseases. As shown in Table 1, a normal sound is attained at 100–200 Hz, and when sensitive, it is detected at a frequency of up to 800 Hz. Adventitious sounds were divided into wheezes and crackles. Wheezes are common in patients with airway obstructions. Typically, COPD and asthma detect breathing at a frequency of 400 Hz. Crackles are divided into subtle sounds according to their pitch and timing. It is characterized by the sound of bubbles bursting discontinuously and is distributed in the frequency band of 60–2000 Hz. The representative diseases include interstitial lung disease (ILD), bronchiectasis, and pneumonia [8,9].

Lung auscultation is considerably influenced by the doctor's

* Corresponding author.
E-mail addresses: youngjin1206@korea.ac.kr (Y. Choi), hclee@korea.ac.kr (H. Lee).

https://doi.org/10.1016/j.bspc.2023.104695
Received 5 August 2022; Received in revised form 21 December 2022; Accepted 11 February 2023
Available online 2 March 2023

Table 1
Characteristics of respiratory sounds: Normal, Wheeze, and Crackle [8,9].

Type	Continuous	Frequency	Pitch	Cause	Disease
Normal	–	100–200 Hz	High (>800 Hz)	–	Asthma, COPD
Wheeze	O	400 Hz	High (>400 Hz)	Airway narrowing, airflow limitation	Asthma, COPD
Crackle	X	60–2000 Hz	Low (<350 Hz)	Explosive opening of small airways (fine crackle) and air bubbles in large bronchi (coarse crackle)	ILD, Bronchiectasis, Pneumonia

experience, patient's condition, external environment, and stethoscope frequency [10]. The frequency of respiratory sounds measured by the stethoscope system is 100–2000 Hz, whereas humans are sensitive only to frequencies of 1000–2000 Hz [11]. Low-quality breath sounds complicate symptom monitoring and diagnosis or lead to misdiagnosis [12]. In addition, it is difficult to evaluate breath patterns during auscultation because a doctor's experience with abnormal sounds is very important. There is a limit to finding abnormal breathing characteristics if the patterns are complex or many. Several systems have been used to solve problems, but these systems have limitations [13,14]. Therefore, this study classified breathing patterns after preprocessing using a bandpass filter to solve the nonlinear pattern problem and misdiagnosis of lung diseases.

Deep learning research that classifies biometric data using an intelligent stethoscope or mobile device to classify lung diseases is being actively developed. Machine learning and artificial intelligence based on deep learning architecture have been applied in several fields, such as cough sound, lung auscultation, lung chest radiography [15], and COVID-19 [16,17]. Artificial intelligence technology can be analyzed and utilized more accurately with a new approach to respiratory sounds [18]. Deep learning can objectively compensate for inaccurate auscultation by clinicians and help in the rapid diagnosis and appropriate treatment of lung diseases [19]. Related works have automatically classified respiratory sounds based on deep learning. As a related sound classification methodology, convolutional neural network (CNN), such as VGGish, L3-Net [20], and transfer learning, is widely used in lung disease classification research. Shi et al. [10] classified lung diseases using transfer learning-based VGGish and BiGRU. The temporal characteristics of respiration were considered using BiGRU. Ponomarchuk et al. [21] classified COVID-19 symptoms by combining VGGish embedding and ensemble models (gradient boosting, lightweight CNN, and logistic regression) using speech, breathing, and coughing signals. Altan et al. [22] proposed a methodology for extracting features by decomposing the time–frequency bands using the Hilbert–Huang transform composed of empirical mode decomposition and Hilbert transform. They attained the highest COPD identification performance by using the proposed deep belief networks. Choi et al. [23] proposed a lightweight CNN-GRU skip connection combining a CNN and BiGRU by collecting respiratory sounds from outpatients and inpatients using Littmann 3200. Through feature stacking, the characteristics of respiratory sounds were emphasized, and a high performance of 92.3 % was confirmed.

The previous study aimed to classify the respiratory sound using original VGGish optimized for audio embedding to reflect the characteristics of the respiratory sound. A performance improvement model reflecting respiratory sound characteristics was proposed using BiGRU, and an ensemble model using machine learning was proposed. However, the types of diseases classified in previous studies (normal, asthma, pneumonia, cough, etc.) were few, and there was a limitation that it was not possible to confirm which characteristics affected classification performance. Moreover, owing to the nature of the algorithm, the possibility of interpreting detailed information regarding the analysis results in the form of a black box is insufficient [18]. This study emphasized the characteristics of respiratory sound by applying the light attention connected module. We analyzed the respiratory sound classification results using Grad-CAM and interpreted the section where abnormal breathing appeared. Related works suggested a smaller number of layers and mentioned weight reduction; we confirmed the advantages of the model by presenting a lightweight model even though it is a deep structure convolution.

Therefore, in this study, we modified VGGish among audio signal-based transfer learning methods to classify lung diseases and suggested the possibility of interpreting the results. A respiratory specialist collected data by measuring the respiratory sounds using a Littmann 3200 stethoscope. The collected data were provided directly after labeling for the normal and five diseases. For preprocessing, the necessary band was obtained through a bandpass filter to remove noises of respiratory sounds. Respiratory information was extracted with the log-Mel spectrogram MFCC features. We classified high performance by applying a light attention connected module to the improved VGGish model proposed. Accordingly, the contributions of this study are as follows.

- We constructed the architecture of a lightweight model by reducing weights and minimizing parameters.
- The light attention connected module emphasized the characteristic information of respiratory sounds and improved the model performance.
- We analyzed and interpreted significant respiratory patterns for the causes of disease classification and the results of attention application. The clinician checked for the symptoms of inspiration or expiration within 5 s of respiration. As a decision-making aid for lung disease diagnosis, it bridges the medical gap.

The remainder of this paper is organized as follows. Section 2 describes previous studies on lung disease classification, attention, and eXplainable artificial intelligence (XAI) using deep learning. Section 3 presents a model that combines the improved VGGish, and attention proposed in this study and describes the data collection, noise removal, and feature extraction methods. Section 4 describes the experimental environment and evaluation indicators, and Section 5 describes the experiment and presents comparisons and analyses of the experimental results. Finally, Section 6 provides the discussion, and Section 7 presents the conclusions, limitations, and prospects of the study.

2. Related work

2.1. Deep learning with CNN

Deep learning using CNNs has made great strides in computer vision. ConvNet is being continuously studied prominently for computer vision and audio processing analysis, and it has recently been widely applied to time series analysis. ConvNet improves the generalization performance of time series analysis through many hidden sides, transfer learning, and significant feature learning of feature activation maps [24]. Studies are being conducted to classify lung diseases using respiratory sounds, which are one of the bio-signals. Lung disease classification research has focused on the collection of publicly available datasets and real-world respiratory sounds, and several attempts have been made to use machine learning and deep learning architectures. Additionally, research on the development of weight reduction is being actively conducted to increase learning efficiency and reduce the parameters of the model

used. Gupta et al. [25] extracted the lung sounds from a Littmann 3200 stethoscope using Gammatonegram and converted them into images. The classification model was presented for the extracted features using the transfer learning of pre-trained AlexNet, GoogleNet, ResNet50, and Inception v3. Asatani et al. [26] presented a classification model using the respiratory dataset from the International Conference on Biomedical and Health Informatics (ICBHI). A short-time Fourier transform (STFT) was applied to the signal to extract spectrograms for time and frequency, and a convolutional recurrent neural network was used. The performance achieved 63 % and 83 % sensitivity and specificity, respectively. Shuvo et al. [27] presented a respiratory disease classification model using a lightweight CNN. A hybrid-scalogram feature extraction method was designed using empirical mode decomposition and continuous wavelet transform, and the accuracy of three and six types was 98.2 % and 98.72 %, respectively. Kranthi Kumar et al. [28] proposed a lightweight CNN using Enhanced-GFCC and Modified-MFCC techniques for SARS-CoV-2/COVID-19 approved at Cambridge University. The accuracy was improved by 4–10 % to 91 % compared to the basic MFCC.

A study on the classification of respiratory sounds related to COPD is also in progress. Altan et al. [29] analyzed COPD lung sounds of 12 channels with five severities of RespiratoryDatabase@TR. COPD is divided into five symptoms (risk, mild, moderate, severe, and very severe). They measured auscultation sounds from the anterior (chest) and posterior (back) and proposed a quantization analysis of the 3D space using three consecutive signal points. They classified COPD with a classification performance of 95.84 %, 93.34 %, and 93.65 % for accuracy, sensitivity, and specificity, respectively. Then, they [30] proposed a methodology for extracting lung sound features by applying 3D second-order difference using a deep extreme learning machine classifier (deep ELM). The performance of the model achieved accuracy, sensitivity, and specificity for 94.31 %, 94.28 %, and 98.76 %, respectively. The limitation of the study was the difficulty in diagnosing because the general severity and relationship between smoking were confused with a small number of datasets.

2.2. Attention

In this study, the performance of the model was improved by emphasizing the characteristic information of the respiratory sound, excluding the specific feature selection step. Attention intensively learns important parts of feature information when training a deep learning model. Representative attention utilizes squeeze-and-excitation networks (SENet), a convolutional block attention module (CBAM), and an ECA-Net. Attention is applied to the multiple layers of the classification model to emphasize and learn feature information. Qayyum et al. [31] classified the COVID X-rays using SENet and depthwise separable convolution. Compared with transfer learning, they showed an accuracy of more than 96.17 %. Dar et al. [13] extracted the main features of respiratory sounds using the bark frequency cepstral coefficient, spectral flux, and spectral centroid. They proposed a hierarchical attention network structure that combined a CNN and LSTM. The model showed an accuracy of 92.3 %. Chen et al. [32] proposed a squeeze-and-excitation CNN for lung cancer classification. Adding only two SE blocks to the classification model improved the performance by 1 %∼2 % compared to the baseline, and the efficiency of attention was demonstrated. However, there was a limitation that more than 85 % of the performance was required to detect benign and malignant tumors. Related studies have used simple attention mechanisms or SE blocks. Therefore, we propose a light attention connected module using ECA-Net.

2.3. Interpretation with Grad-CAM

XAI is an explanatory artificial intelligence model that can help interpret the black box. Most previous XAI studies have applied Grad-CAM to chest X-rays for visual analysis and interpretation. Haghanifar et al. [33] presented a COVID-CXNet model to classify published COVID-19 chest radiography images. The model obtained an accuracy of 87.88 %, and the classification results for COVID-19 and pneumonia were visually confirmed using Grad-CAM. Zhang et al. [34] converted the voice data of the 2019 TAU Urban Acoustic Scenes into a Mel spectrogram, classified the voice using ResNet20, and applied Grad-CAM to interpret the results. CAM was applied and analyzed to the Mel spectrogram and MFCC, which converts 1D-based sound into a 2D image. The characteristics of the Mel spectrogram were efficiently analyzed to visualize the CNN. XAI utilizing Grad-CAM has been extensively studied only in X-rays; studies related to lung sounds are incomplete and require more research. This study presents a classification analysis of respiration information by applying Grad-CAM to a log-Mel spectrogram MFCC feature information for respiratory sounds.

3. Materials and methods

This section describes the architecture of data collection and preprocessing, feature extraction, attention, the proposed method, and interpretation based on Fig. 1.

3.1. Data - clinical dataset

Respiratory sounds were measured in outpatients and inpatients (average age, 66 years; height, 160 cm; weight, 60.5 kg) aged $\geq$ 19 years who visited the respiratory medicine department. The clinician used chest X-rays and CTs to label the patient's lung disease. The consent of patients and the approval of the Institutional Review Board of the Human Research Ethics Committee were obtained (Approval No. KC20ONSI0774). The respiration was measured at four sites by a respiratory specialist using a Littman 3200 stethoscope. The specialist measured the respiratory sounds within 1 min. The respiratory sounds were measured from the right and left sides, respectively, in the upper and lower parts of the anterior region (chest) and the posterior region (back). The data collected in this study were classified into normal, crackle, and wheeze, and the respiratory sounds were measured for patients who satisfied the conditions in Table 2. As shown in Table 2, wheezing consists of COPD and asthma. Auscultation of symptoms such as cough, sputum, and shortness of breath appear. The crackle sounds were collected from the patients with the disease using chest X-rays and CTs. The respiratory sounds for ILD, bronchiectasis, and pneumonia were collected by auscultating crackles at the lesion site.

3.2. Preprocessing & feature extraction (stage 1)

3.2.1. Segmentation of respiratory sounds

When measuring respiration, the protocol was set at intervals of 2–3 s for inhalation and 1–2 s for exhalation. We collected experts' opinions and constructed 1,021 datasets by dividing a total of 126 respiratory sounds within 1 min into 5 s. As listed in Table 3, respiratory sounds were used to train and evaluate the model. The sampling rate of respiratory sound was sampled at 4000 Hz. A sampling at 4000 Hz also had the effect of removing ambient noise [22].

3.2.2. Bandpass filter

Respiratory sounds generate noise during measurement, depending on the treatment environment and time. In this study, noises, such as skin friction generated during measurement, were removed using a bandpass filter, as shown in Fig. 2. We applied the 5th-order Butterworth filter and the frequency band of 250–1800 Hz to the respiratory sound and verified the experimental performance by comparing each pass filter pretreatment in the experimental results in Section 5.

3.2.3. Log-Mel spectrogram MFCC

This study converted 1D respiration signals into 2D time–frequency log-Mel spectrogram mel frequency cepstral coefficient (MFCC) for

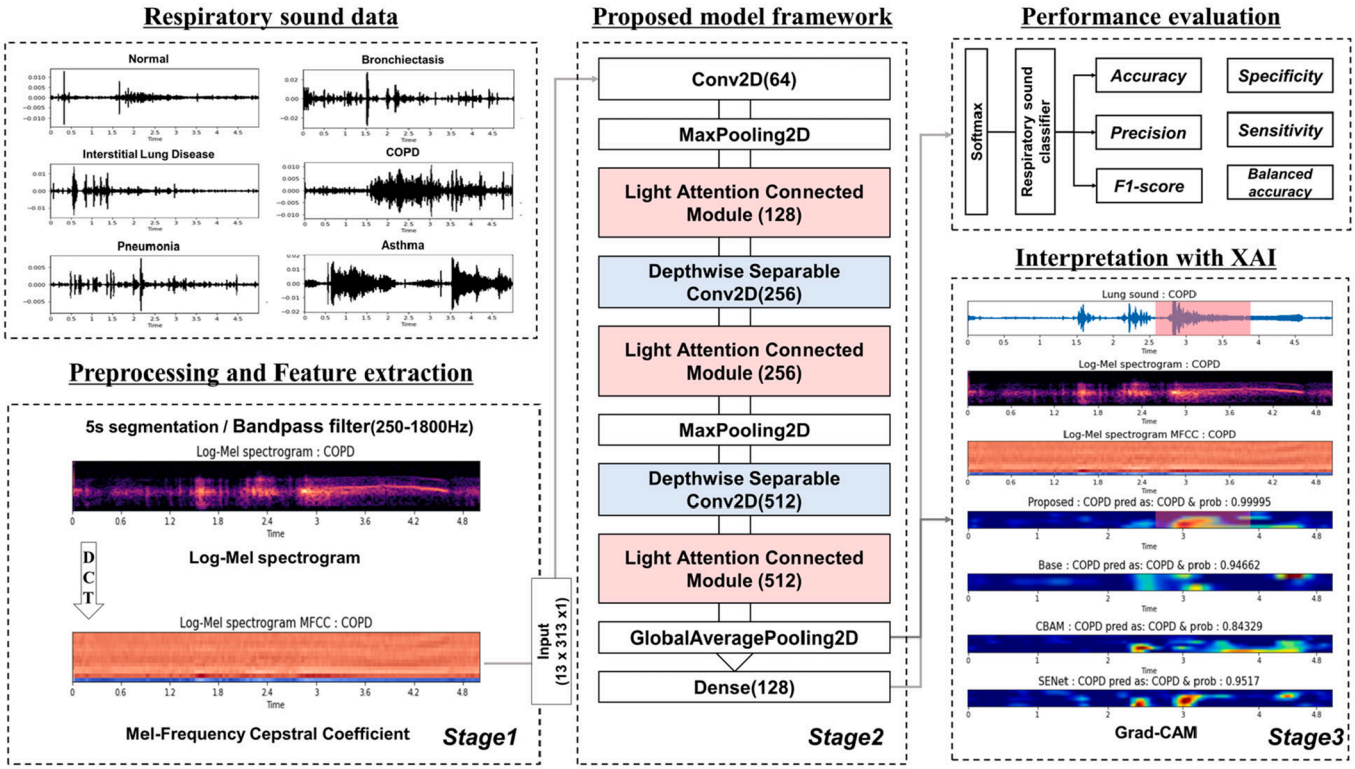


Fig. 1. Architecture of the proposed model.

Table 2
Information of respiratory sound and measurement conditions.

Symptom	Normal	Wheeze		Crackle		
Disease	**Normal**	**Asthma**	**COPD**	**ILD**	**Pneumonia**	**Bronchiectasis**
Condition	Normal breathing	New or worsening symptoms, such as cough, phlegm, or shortness of breath		Auscultation of crackles at the lesion site	New lesions on chest X-rays or CT one week before enrollment Auscultation of crackles at the lesion site	Auscultation of crackles in the area of bronchiectasis on chest imaging
Number of data	12	23	20	26	20	25

Table 3
Segmentation of the clinical dataset.

Clinical Dataset		
Symptom	**Disease**	**Total**
Normal	Normal	114
Wheeze	COPD	160
	Asthma	300
Crackle	Pneumonia	200
	Bronchiectasis	160
	Interstitial lung disease	203
Total		1021

feature extraction and classification model input. Log-Mel spectrogram was used as an advantage of time–frequency characteristics that reflected respiratory sounds of low and high frequencies compared to STFT, wavelet transform, and Stockwell transform [35]. Log-Mel spectrogram is the information obtained by converting the spectrogram to the Mel-scale and shows good performance in audio and respiratory sound classification [6]. The MFCC is a method for extracting the unit energy of an input signal. The energy coefficient of the characteristic frequency section was obtained by performing a discrete cosine transform (DCT) on the log-Mel spectrogram. Therefore, in this study, the log-Mel spectrogram MFCC [23] was used as an input for the proposed model. The MFCC generation process shown in Fig. 2 is as follows.

1) Respiratory sounds are continuously input and passed through a window filter ($w(n)$) that divides the signal into short intervals. N is the length of the window function and Eq. (1) is as follows:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) 0 \le n \le N \tag{1}$$

2) After passing through the window filter, a spectrogram of the signal-frequency band was obtained using STFT. The log-Mel spectrogram is a spectrogram of the time–frequency band in which log transformation is performed after calculating the spectrogram using the Mel filter bank in Eq. (2).

$$f_{mel} = 2595 log_{10}(1 + f/700) \tag{2}$$

3) MFCC was converted to a log-Mel spectrogram through DCT [36]. A total of 313 frames overlapped by 75 %, each with a window length

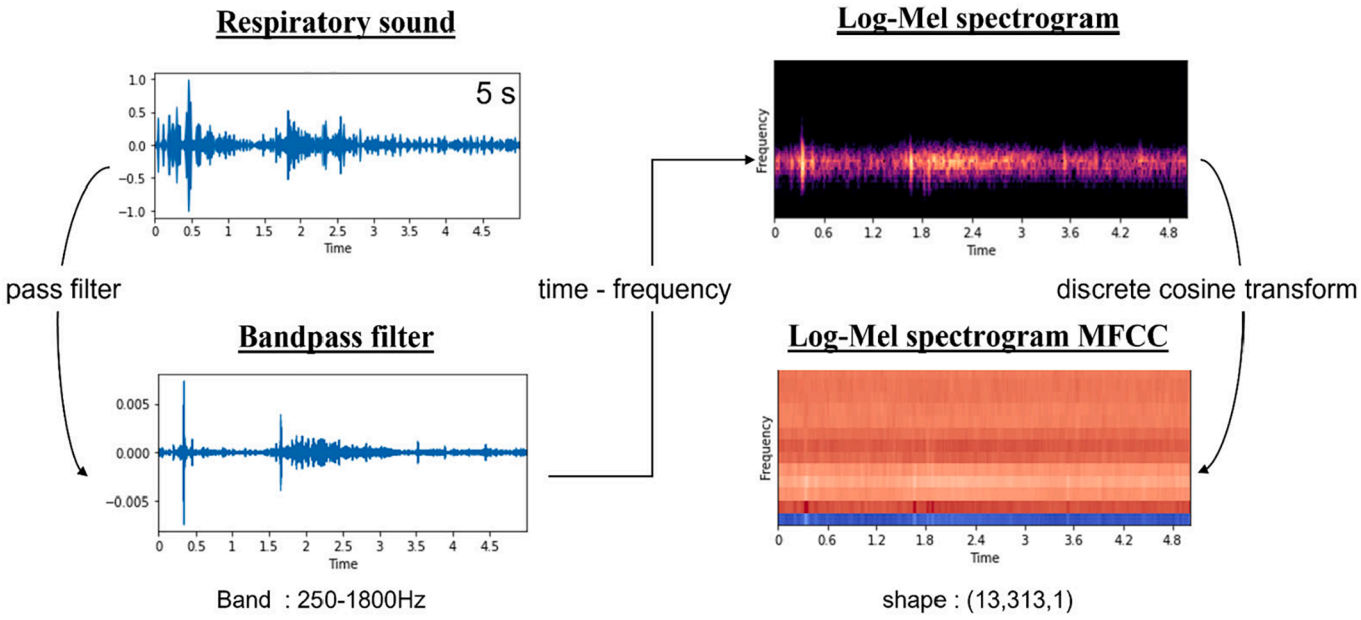


Fig. 2. Signal to log-Mel spectrogram MFCC.

of 64 ms. The MFCC coefficient, window size, and hop length were set to 13, 256, and 64, respectively in Table 4.

3.3. Attention modules

3.3.1. Squeeze-and-excitation networks (SENet)

SENet is a module that improves performance after calculating the weights between channels using squeeze and excitation. SE block compresses and converts input ($X \in \mathbb{R}^{H' \times W' \times C'}$) into the feature map ($U \in \mathbb{R}^{H \times W \times C}$). Squeeze ($z_c$) transforms the $H \times W \times C$ feature map into one-dimensional (1)$1 \times 1 \times C$ for each channel through a global average pooling operation in Eq. (3). Excitation (S) is weighted after $1 \times 1 \times C$ normalization generated from squeeze Eq. (4). This activates the squeeze with the ReLU function after calculating the weight ($W_1 \in \mathbb{R}^{(\frac{C}{r}) \times C}$) and the fully connected product. After the activated value is multiplied by the weight ($W_2 \in \mathbb{R}^{C \times (\frac{C}{r})}$) and the fully connected product, it is activated with a sigmoid function to obtain the excitation [37]. Since the hyperparameter does not increase significantly, there is an advantage that the amount of calculation of the parameter is small. However, although dimension reduction is performed in the fully connected layer, as shown in Fig. 3(a), it is reduced by a ratio equal to the reduction ratio, which breaks the direct correspondence between the channels and weights.

(1) Squeeze:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j), U = [u_1, u_2, \cdots, u_c] \tag{3}$$

(2) Excitation:

$$S = sigmoid(g(z, W)) = sigmoid(W_2 ReLU(W_1 Z)) \tag{4}$$

Table 4
Parameters of the log-Mel spectrogram MFCC.

Parameter	Sampling rate	FFT Window size	Hop length	Number of Mel bins	Number of MFCC
Value	4000 Hz	256	64	64	13

3.3.2. Efficient channel attention (ECA-Net)

Wang et al. proposed ECA-Net as an attention module to improve the problem occurring in SENet [38]. Channel attention is used similarly to reduce the complexity of the model. Fig. 3(b) and Eq. (5) indicate that ECA-Net (eca_ω) uses GAP ($g(X)$) and $conv1D$ with kernel size (k) as a fully connected layer to lower complexity instead of dimension reduction, unlike SENet.

$$g(X) = \frac{1}{H \times W} \sum_{i=1,j=1}^{H,W} X_{i,j} \tag{5}$$

$$eca_\omega = \sigma(conv1D_k(g(X)))$$

ECA-Net has the advantage of considering local cross-channel interactions and designing a lightweight model. To reduce the complexity of cross-channel interaction and maintain performance, it is necessary to set appropriate variables. The mapping function (ϕ) of the channel (C), including kernel size (k) and interaction, is given by Eq. (6).

$$C = \phi(k) = 2^{(\gamma - k - b)} \tag{6}$$

The mapping function of k and C has limitations when characterizing the coverage of the interaction as a linear function. C is a nonlinear function; when converted into a function for C, it is given by Eq. (7). $|t|_{odd}$ is the nearest odd number of t, and γ and b are set to 2 and 1, respectively, in all experiments.

$$k = \varphi(C) = \left| \frac{log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{7}$$

3.4. Proposed models (stage 2)

3.4.1. Modified VGGish

In this study, a respiratory sound classification model was designed based on VGGish, which is widely applied to sound classification. VGGish is a deep CNN proposed by Hershey et al. [39] based on the VGGNet structure during transfer learning in computer vision. VGGish includes a deep audio embedding mode and is the proposed method for classifying audio from YouTube videos. Pre-trained VGGish is often used for audio classification [40]. A characteristic of the model structure is that several feature extractions are performed using the four block structures combined by convolution and max pooling. It also includes

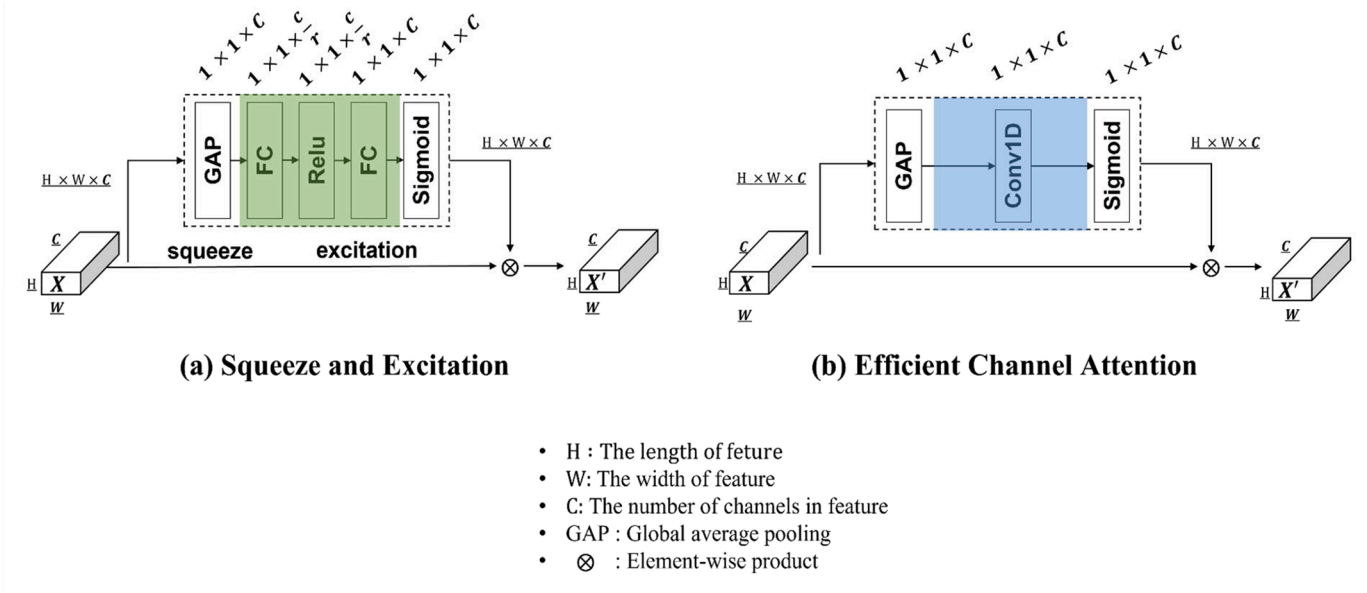


Fig. 3. Attention modules: (a) SENet and (b) ECA-Net.

the structure of a fully connected layer that performs as a classifier [41]. In this study, to improve model performance and apply the XAI methodology, the flattened layer and max pooling of VGGish were converted to global average pooling. This reduces the number of parameters that require model calculation and effectively prevents overfitting. In addition, the baseline is improved by applying global average pooing, LeakyReLU, batch normalization, and spatial dropout. A depthwise separable convolution was used to improve the model with deeper layers and reduce the amount of computation. The structure is a convolution constructed by merging a depthwise structure that separates the feature map for each channel and performs an operation and a 1 × 1 pointwise convolution structure that merges multiple channels into one new channel [23].

3.4.2. Light attention connected module

We proposed a structure that combines ECA-Net and depthwise separable convolution for the light attention connected module (LACM). This partly cites the block structure combining the SENet module of the inverted residual block in the MobileNetV3 [42]. Pseudo codes are summarized in Table 5. The input features ($Input_{logmelmfcc}$) are passed through a depthwise separable convolution. By inputting the obtained X_{depth} value into ECA-Net, attention is calculated. (Algorithm 1., 1–2) ECA-Net extracts attention through the input (X_{depth}) and output (X') product by global average pooling, conv1d, and sigmoid. (Algorithm 1., 3–7) The extracted value concatenates attention by combining $Input_{logmelmfcc}$ and X'_{eca}. As shown in Fig. 4(d), it is a connected structure and is transmitted to prevent loss of function and maintain information.

Table 5
Pseudo code for light attention connected module.

Algorithm 1. Pseudo code for light attention connected module.	
Input: Log-Mel Spectrogram MFCC $Input_{logmelmfcc}$	
Output: Attention network $LACM_{Feature}$	
1	*Lightweight Attention Network*
2	$X_{depth} \leftarrow$ *Depthwise Separable Conv2D* $(Input_{logmelmfcc})$
3	$X'_{eca} \leftarrow ECANet(X_{depth})$
4	$X' \leftarrow Globalaveragepooling(X_{depth})$
5	$X' \leftarrow Conv1D(X')$
6	$X' \leftarrow Sigmoid(X')$
7	$X'_{eca} \leftarrow$ *Multiply* $([X', X_{depth}])$
8	$LACM_{Feature} =$ *Concatenate* $(Input_{logmelmfcc}, X'_{eca})$

The performance of the proposed model was compared with the results obtained using the SENet and CBAM in Section 5.

3.4.3. Grad-CAM (stage 3)

The class activation map (CAM) is an explainable AI model proposed by Zhou et al. [43] to interpret model prediction and classification results. This weakly supervised learning method computes weights between convolution and target classes. w_k^c is the k_{th} feature map that predicts the c class, and the CAM calculation is expressed as follows:

$$L^c_{CAM} = \sum_k w_k^c f^k \tag{8}$$

Since the CAM focuses on pixel label information rather than training image information, it visually interprets the feature information of the last layer. However, because the information is lost in the flattened layer, Selvaraju proposed Grad-CAM [44]. It is a visualization methodology that generalizes CAM for various CNN architectures by calculating the weight of the feature map, which can explain the class as a gradient. The calculation process of Grad-CAM, given by Eq. (9), is expressed as follows:

1. Calculate a_k^c that passed global average pooling ($\frac{1}{z}\sum_i\sum_j$) by differentiating the k channel feature map ($A^k \in \mathbb{R}^{\mu^*\nu}$) for the target class with the predicted class score (y^c) for class $c(\frac{\partial y^c}{\partial A_{ij}^k})$. Eq. (10) is as follows:

$$L^c_{Grad-CAM} \in \mathbb{R}^{\mu^*\nu} \tag{9}$$

$$a_k^c = \frac{1}{z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{10}$$

2. Multiply a_k^c using the feature map (A^k) to calculate the weight obtained from the activation map and apply the ReLU. The ReLU is used to identify only the significant features that positively affect class c. Eq. (11) is as follows:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k a_k^c A^k\right) \tag{11}$$

Therefore, in Subsection 5.4, we analyzed the disease group

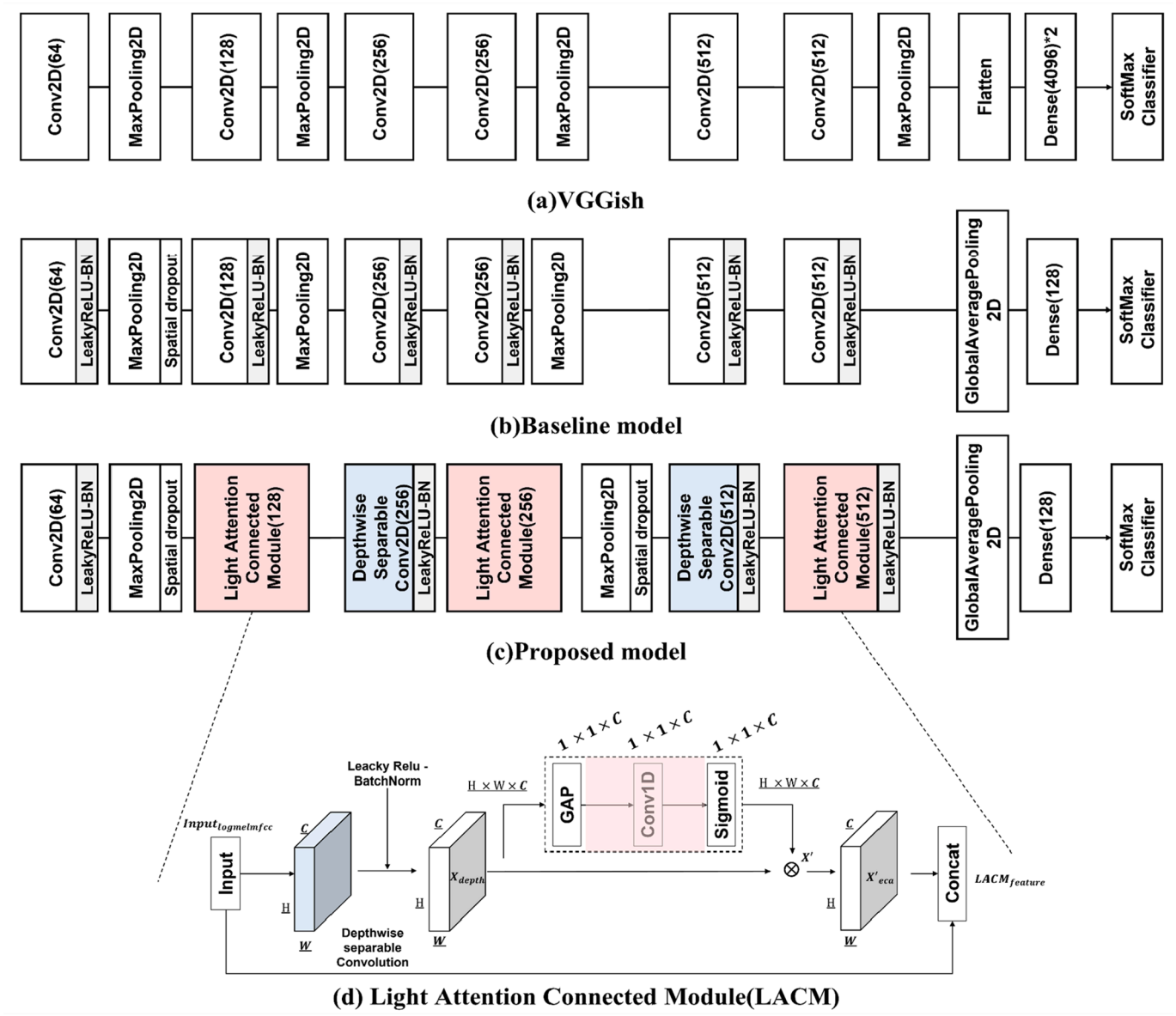


Fig. 4. Models with (a) VGGish, (b) baseline, (c) proposed model, and (d) light attention connected module.

classification results using Grad-CAM and confirmed the effect of attention.

4. Experiments

4.1. Hyperparameter settings

The experimental setup was implemented using Python 3.8 and TensorFlow 2.4.0 on a computer with an AMD Ryzen 7 2700X (CPU), 64 GB RAM, and GeForce 1080TI (GPU). The loss and optimization functions used cross entropy and Adam optimizer, respectively. The hyperparameters in Table 6 were set to a batch size of 8, epoch 500, and learning rate 3e-4-3e-5. The model was improved by adjusting the learning rate using a learning rate scheduler. For the validation of the dataset, 5-fold cross validation was used. All comparison models were applied equally to the hyperparameters.

Table 6
Hyperparameter settings.

Parameter	Value
Spatial dropout rate	0.2
Learning rate	3e-4-3e-5 (ReduceLR On Plateau)
Optimizer	Adam
Batch size	8
Epoch	500
Cross validation	5

4.2. Model structure

We constructed the classification model using the structure in Table 7. The proposed model was trained using an input of $13 \times 313 \times 1$ and parameters of 802,194.

4.3. Evaluations

For the model evaluation, five performance indicators were used: accuracy, precision, sensitivity (12), specificity (13), f1-score, and balanced accuracy (14) [45].

$$Sensitivity(SE) = \frac{TP}{TP + FN} \tag{12}$$

Table 7
Details of the model structure.

Layer	Output size	Params
Input	(None, 13, 313, 1)	–
Conv2D (3 × 3, 64)	(None, 13, 313, 64)	640
Max pooling / Spatial dropout (rate = 0.2)	(None, 13, 156, 64)	256
LACM 1(3 × 3, 128)	(None, 13, 156, 192)	9,412
Depthwise separable conv2D (3 × 3, 256)	(None, 13, 156, 256)	52,160
LACM 2 (3 × 3, 256)	(None, 13, 156, 512)	69,124
Max pooling / Spatial dropout (rate = 0.2)	(None, 6, 78, 512)	–
Depthwise separable conv2D (3 × 3, 512)	(None, 6, 78, 512)	269,312
LACM 3 (3 × 3, 512)	(None, 6, 78, 1024)	269,316
Global average pooling	(None, 1024)	–
Dense layer	(None, 128)	131,200
SoftMax	(None, 6)	774
Total number of parameters		**802,194**

$$Specificity(SP) = \frac{TN}{TN + FP} \tag{13}$$

$$Balancedaccuracy = \frac{Sensitivity + Specificity}{2} \tag{14}$$

5. Results

5.1. Results of the baseline and proposed models

Table 8 shows the performance results of the baseline and proposed models. As shown in Fig. 4(c), the proposed model was designed as a lightweight model using the LACM of ECA-Net and depthwise separable convolution in the baseline. The experiment was conducted using the experiment 5-fold cross validation, and the model performance was the average (±SD) value of cross validation. The proposed model showed high performance, with an accuracy of 92.56 %, precision of 92.81 %, sensitivity of 92.22 %, specificity of 98.50 %, f1-score of 92.29 %, and balanced accuracy of 95.4 %. In particular, sensitivity and specificity are widely used as the most important indicators in disease classification. Sensitivity refers to predicting a disease positively, and specificity is the prediction ratio of the absence of a disease when there is no actual disease. The better the performance evaluation, the better the reliability of the model. Compared with the baseline, the performance of the proposed model improved in terms of the indicators by 3.33 %, 3.45 %, 3.21 %, 0.66 %, 3.32 %, and 2 %, respectively.

We also analyzed the inference time of the results by citing Choe et al. [46]. The average inference time for 50 times was an average of 387 MS, a minimum of 606 MS, and a maximum of 960 MS, respectively. The computation time was 21 MS and 30 MS for the baseline and proposed models, respectively, for one training based on 500 iterations. The inference time was 508 MS and 706 MS for the baseline and the proposed models, respectively. Although there was a 226 MS (0.3 s) difference from the baseline, the reasoning speed was faster than 1 s. The number of parameters was reduced by 80 % by applying depthwise separable convolution, and the lightweight effect was confirmed. When considering flops, the proposed model showed a lower amount of computation.

Table 8
Comparison of the baseline and proposed models (%).

Evaluation	Baseline model	Proposed model	Improvement
Accuracy	89.23(±2.62)	92.56(±1.4)	3.33
Precision	89.36(±2.86)	92.81(±0.86)	3.45
Sensitivity	89.01(±2.54)	92.22(±1.65)	3.21
Specificity	97.84(±0.52)	98.5(±0.3)	0.66
F1-score	88.97(±2.63)	92.29(±1.33)	3.32
Balanced accuracy	93.4(±1.53)	95.4(±0.97)	2
Parameters	4,573,062	802,194	−80
Computation time (MS)	21	30	–
Inference time (MS)	508	706	–
FLOPS (billion)	1.9	1.0	

5.2. Confusion matrix of baseline and proposed models

Table 9 and Fig. 5 present the confusion matrix result for the baseline and proposed models. The result by the proposed model improved the score for all diseases. In particular, in the case of pneumonia with symptoms similar to COVID-19, pneumonia was partially confused with the normal and ILD symptoms at the baseline. However, the proposed model was classified as having high performance, as it could distinguish normal respiratory sounds successfully. Similarly, by classifying bronchiectasis, normal, and pneumonia diseases, the possibility of providing information on the early diagnosis of crackle symptoms to medical staff was confirmed based on the results.

5.3. Interpretation of lung disease with XAI (Grad-CAM)

We presented the visualization results of Grad-CAM to explain the evaluation of the proposed model. This study used Grad-CAM as the analysis model because there was no change in the model structure and no loss in the classification result. We also tried to ease the visualization with the log-Mel spectrogram and MFCC. The experimental results confirmed that information was concentrated in the activation map of the feature when attention was used. For each model, it was possible to visually identify the characteristic points at which diseases were classified.

At the time of data collection, explanatory power was added to the XAI results by collecting respiratory sounds, including the positions to the right and left sides of the anterior (chest) and posterior (back), respectively, in Fig. 6. Specialists measured it at four sites. We could utilize the information on labeling the chest position for respiratory sounds with lung disease. Practically, we have received opinions from specialists that the wheezing sounds of COPD and asthma can be heard in both lung fields and that the crackles can be different at the locations of lung lesions. Therefore, we have incorporated expert opinion into our interpretation.

As shown in Fig. 7(a), we can be confirmed that a constant respiratory sound is maintained and intensively learned. Normal sounds are heard during inspiration and early expiration [48]. This section corresponds to one cycle of inhalation and exhalation and can be considered a section of normal sound. Clinicians directly identified the intervals for disease judgment.

In the case of pneumonia, the phenomenon of pitch appeared directly in Grad-CAM and was compared with normal. We intensively visualized the bubble bursting in the breathing sound in the lower right posterior chest (RLL), and the predictive power was 0.999. Bronchiectasis and ILD at the right anterior chest (RUL) location were visually confirmed by bubble bursting or rough crackling. The clinician determined that the respiratory sounds were heard in the second area (exhalation) of two diseases but confirmed that an accurate diagnosis was needed using X-rays or CTs. CBAM and SENet demonstrated the accuracy of the proposed model by showing that Bronchiectasis and ILD symptoms were incorrectly predicted as pneumonia. CBAM and SENet predicted more than 58 % probability, and both mispredicted other diseases, as shown in Fig. 7(b).

Failure to diagnose wheezing early can lead to misdiagnosis or disease progression, which can have serious consequences [48]. In COPD and asthma, air blows occur owing to the narrowing of the airways, and it has been confirmed that the phenomenon mainly occurs during exhalation, characterized by a whistling sound. COPD showed abnormal symptoms in the left anterior chest (LUL), and asthma showed abnormal symptoms in the right posterior chest (RLL). Clinicians detected respiratory sounds similar to cough or dyspnea patterns, as shown in Fig. 7 (c). They confirmed that both diseases had high-pitched asthma and COPD symptoms. Therefore, by providing interpretive power in the

Table 9
Result of the baseline and proposed models for the confusion matrix score (%).

Baseline model					Proposed model				
Class 6	**Precision**	**SE**	**SP**	**F1-score**	**Class 6**	**Precision**	**SE**	**SP**	**F1-score**
Bronchiectasis	86.04	85.5	96.46	85.4	Bronchiectasis	89.74	91.5	97.32	90.43
COPD	91.82	83.75	98.61	87.5	COPD	96.27	87.5	99.3	91.4
ILD	92.34	94.58	98.04	93.44	ILD	93.5	98.02	98.29	95.7
Normal	86.54	87.71	98.24	86.88	Normal	90.15	91.26	98.68	90.39
Pneumonia	85.59	91.25	96.99	88.13	Pneumonia	90.44	89.38	98.14	89.65
Asthma	93.82	91.26	98.69	92.5	Asthma	96.78	95.64	99.28	96.18
Average	**89.36**	**89.01**	**97.84**	**88.97**	**Average**	**92.81**	**92.22**	**98.5**	**92.29**

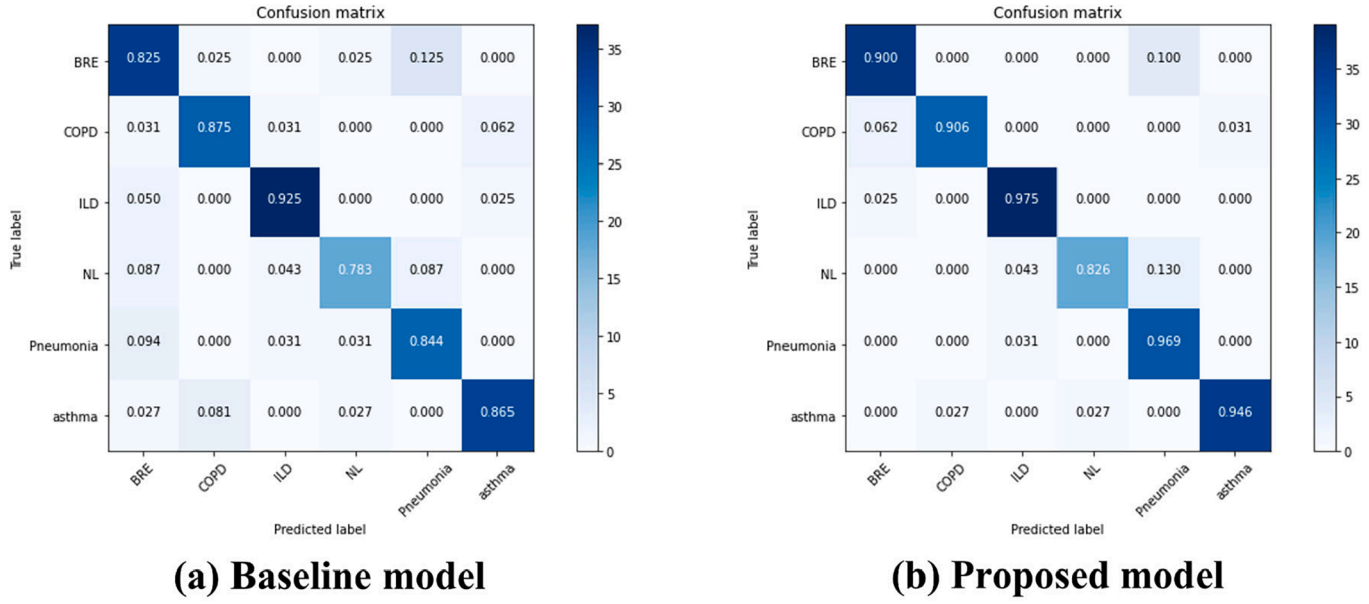


Fig. 5. Confusion matrix with (a) baseline and (b) proposed model.

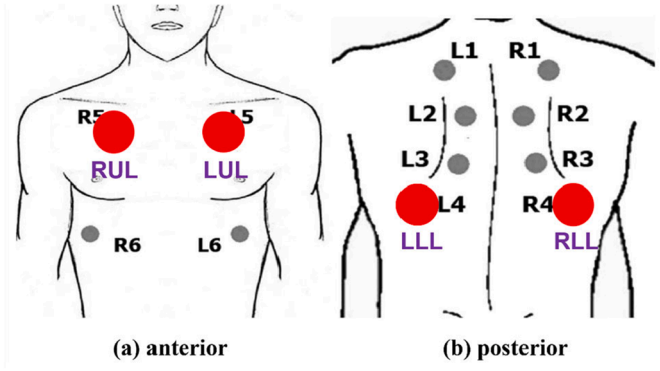


Fig. 6. Respiration measurement location: Both left and right of (a) Anterior upper and (b) Posterior lower [47].

model, COPD, a significant cause of death at an early stage, can be diagnosed early.

5.4. Comparison of attention models

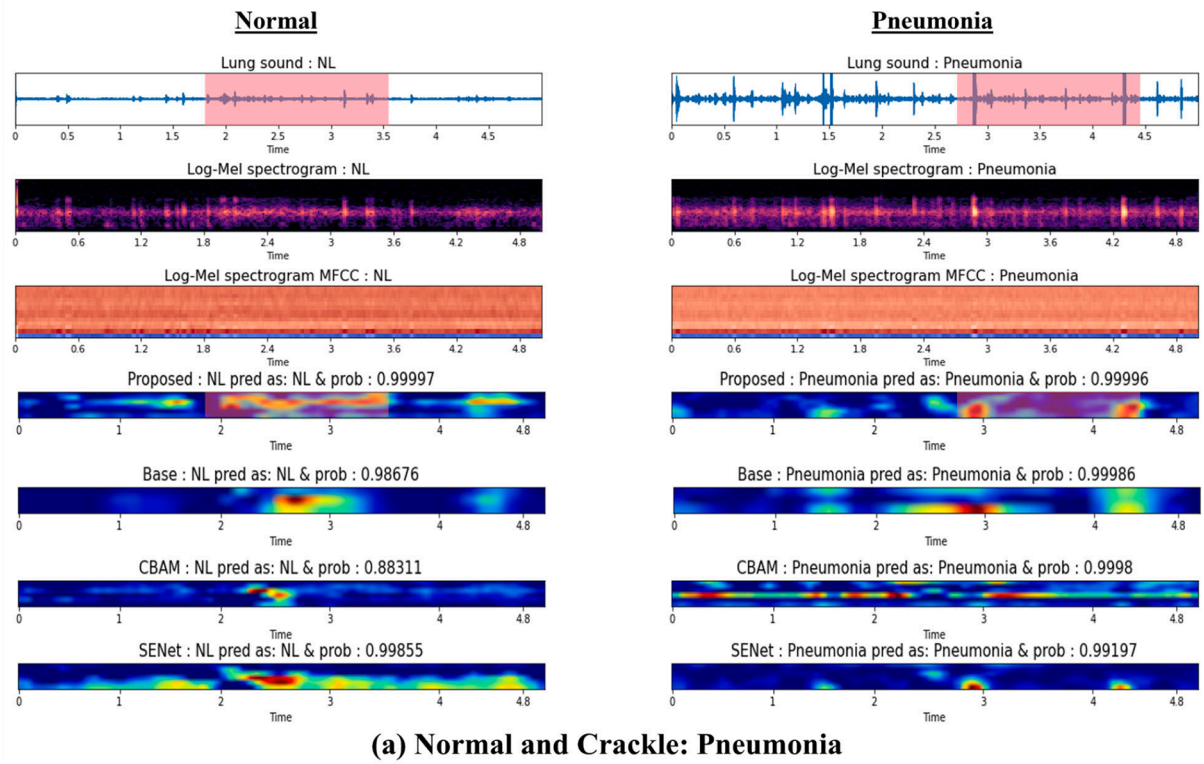
Table 10 compares the experimental results with those of the proposed model's CBAM, SENet, and ECA-Net. We compared and analyzed the model's performance according to attention. As shown in Table 10, the no-attention model achieved an accuracy of 89.81 %, precision of 90.01 %, sensitivity of 89.42 %, specificity of 97.95 %, f1-score of 89.52 %, and balanced accuracy of 93.38 %. The CBAM is an attention module that emphasizes features using channel and spatial attention [49]. The results using CBAM showed an accuracy of 78.05 %, precision of 78.20 %, sensitivity of 77.16 %, specificity of 95.58 %, f1-score of 77.29 %, and balanced accuracy of 86.37 %. CBAM performed worse than the baseline; therefore, there was no effect on attention. The results using SENet showed an accuracy of 90.21 %, precision of 90.17 %, sensitivity of 89.81 %, specificity of 98.03 %, f1-score of 89.86 %, and balanced accuracy of 93.92 %. CBAM and SENet either mispredicted respiratory sounds or had poor predictive power, as shown in Fig. 7.

We compared the model to a previous study using the same Littmann 3200 data set. Choi et al. [23] emphasized the respiratory sound information using feature stacking, but this study emphasized the features of feature information using LACM based on ECA-Net. As shown in Fig. 8, compared to the previous study, the accuracy was 0.3 %, precision 0.81 %, sensitivity 0.11 %, and f1-score improved by 0.4 %. We have confirmed the model's strength by reducing standard deviations compared to previous studies.

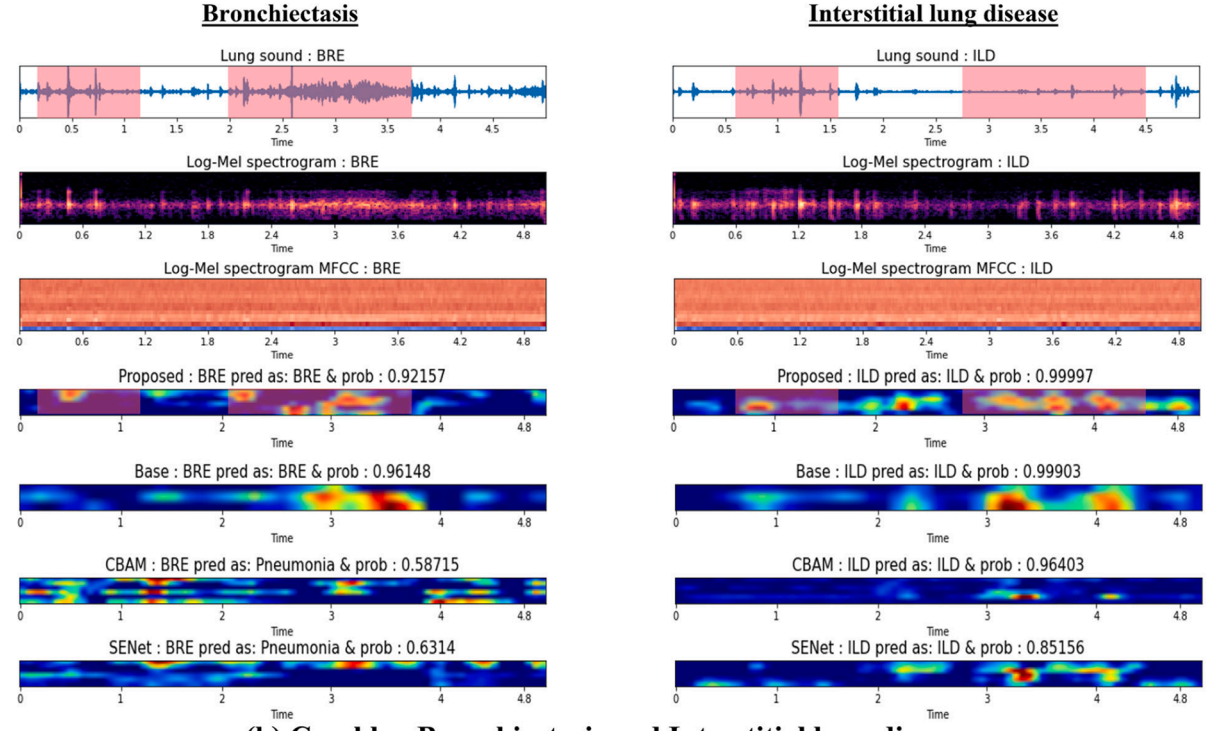
5.5. Comparison of pass filter and augmentation

5.5.1. Result of pass filter

Table 11 compares the performance of the models according to noise removal. The bandpass, lowpass, and highpass filters were analyzed, and Fig. 9 shows the result of applying the filters. For the lowpass filter, respiratory sounds in the 250 Hz or lower band were used, and for the highpass filter, 250 Hz or higher was used. Accordingly, the lowpass filter obtained an accuracy of 80.8 %, precision of 80.41 %, sensitivity of



(a) Normal and Crackle: Pneumonia



(b) Crackle : Bronchiectasis and Interstitial lung disease

Fig. 7. Interpretation of disease with Grad-CAM (XAI): (a) Normal and crackle: pneumonia, (b) Crackle: bronchiectasis and interstitial lung disease, and (c) Wheeze: COPD and asthma.

79.83 %, specificity of 96.15 %, f1-score of 79.87 %, and balanced accuracy of 87.99 %. The highpass filter showed an accuracy of 91.28 %, precision of 91.67 %, sensitivity of 90.91 %, specificity of 98.24 %, f1-score of 91.07 %, and balanced accuracy of 94.57 %. We achieved the highest performance, and the effect of pretreatment on the required band was confirmed. The required band was verified experimentally at 250–1800 Hz.

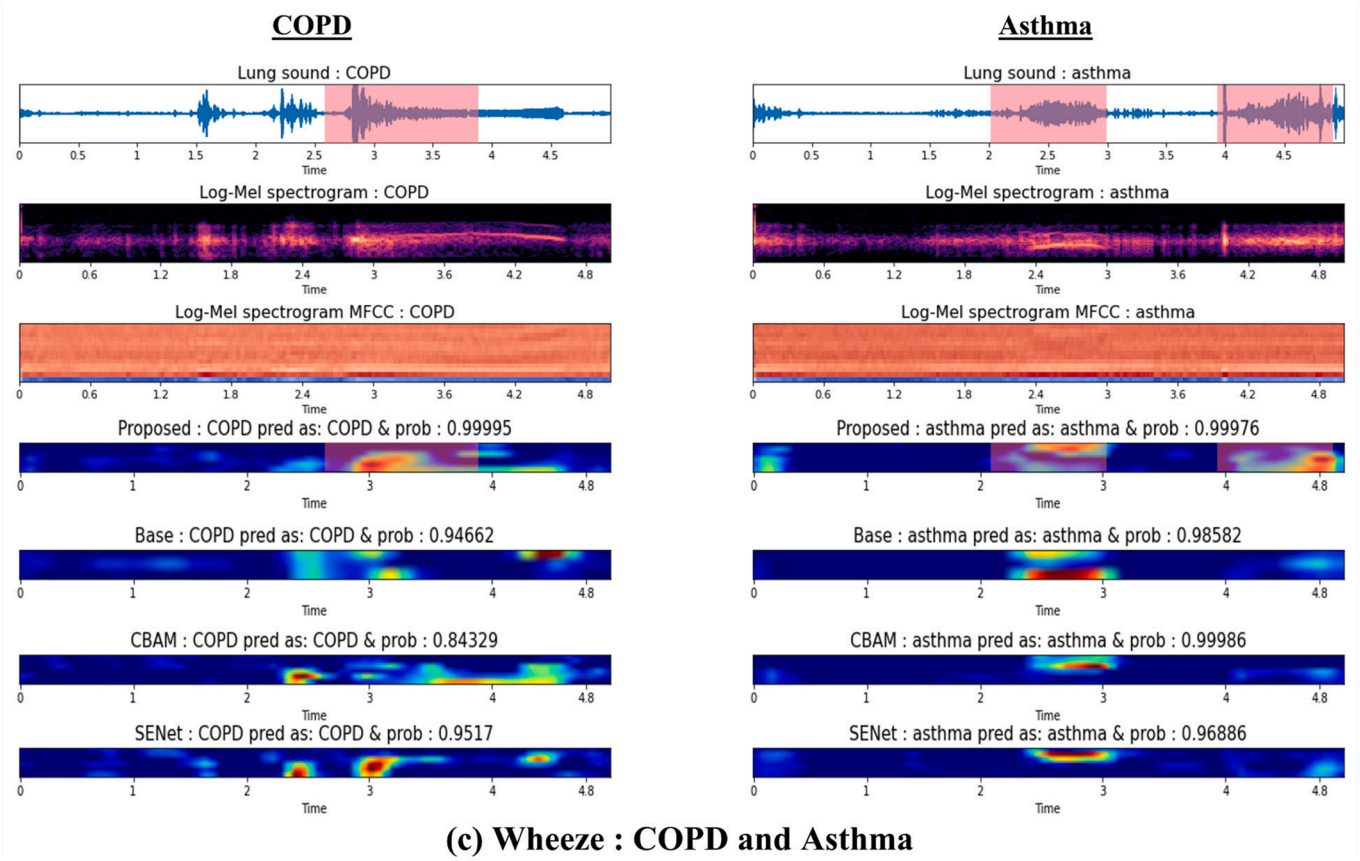


Fig. 7. (*continued*).

Table 10
Comparison of attention with No-attention, CBAM, SENet, and proposed model (%).

Attention	Accuracy	Precision	Sensitivity	Specificity	F1-score	Balanced accuracy
No attention	89.81(±2.64)	90.01(±2.59)	89.42(±3.02)	97.95(±0.53)	89.52(±2.85)	93.69(±1.77)
CBAM	78.05(±3.66)	78.2(±4.17)	77.16(±4.25)	95.58(±0.74)	77.29(±4.2)	86.37(±2.49)
SENet	90.21(±2.73)	90.17(±2.92)	89.81(±2.86)	98.03(±0.54)	89.86(±2.9)	93.92(±1.7)
Proposed model	**92.56(±1.4)**	**92.81(±0.86)**	**92.22(±1.65)**	**98.5(±0.3)**	**92.29(±1.33)**	**95.36(±0.97)**
Choi et al. [23]	92.3(±2.5)	92.0(±2.6)	92.1(±2.6)	98.5(±0.5)	91. 9(±2.6)	95.3(±1.55)

5.5.2. *Result of augmentations*

Table 12 and Fig. 10 present the results of applying augmentation of the time stretch and pitch. Time stretch is a method of increasing the data size by horizontally expanding the data. When converting a 1D respiratory sound into a 2D spectrogram, the speed is adjusted faster and slower without losing information. Pitch is a methodology in which the pitch of a sound is changed. Two types of time stretch and pitch were considered [36], and augmentation effectively improved the class imbalance in model performance [8]. However, the performance of the measurement data collected in this study decreased when the augmentation was applied. According to the clinician's advice, the opinion that the popping sound disappeared when listening to the auscultation sound was confirmed experimentally. The possibility of lung disease classification was demonstrated with an accuracy of 89.81 %.

5.6. *Littmann stethoscope data utilization with data in brief*

We compared and analyzed the respiratory sound data from [50] to validate the lung disease classification model. The data were used to utilize the respiratory sounds measured by the same device (Littmann 3200). The dataset was extracted from the stethoscope's memory using stethoscope software, and the real lung sounds from Middle Eastern participants were recorded. The subjects were aged 21–90 years (mean ± SD: 50.5, ± SD: 19.4). The dataset included sounds from seven diseases (asthma, heart failure, pneumonia, bronchitis, pleural effusion, pulmonary fibrosis, and COPD) and normal sounds. As shown in Table 13, the open dataset used in this experiment was respiratory sound data for asthma, pneumonia, bronchitis, pulmonary fibrosis, and chronic obstructive disease.

We summarized the performance results in comparison with those in [51] and [52] in Table 14. The methodology proposed achieved high performance with an accuracy of 86.37 % (± SD 4.02). In [51], the Mahalanobis distribution and ResNet were used, and in [52], the features were extracted through empirical wavelet transformation, and the respiratory sounds were classified using a light gradient boosting machine (LGBM). The performance of the proposed methodology was the highest in the experimental results, and the effects of attention were confirmed. Moreover, Soni et al. [53] proposed a contrastive learning classification methodology using the dataset as an external validation set. Data was useful enough to be used for model validation.

6. Discussion

We devised a lung disease classification using a modified VGGish and

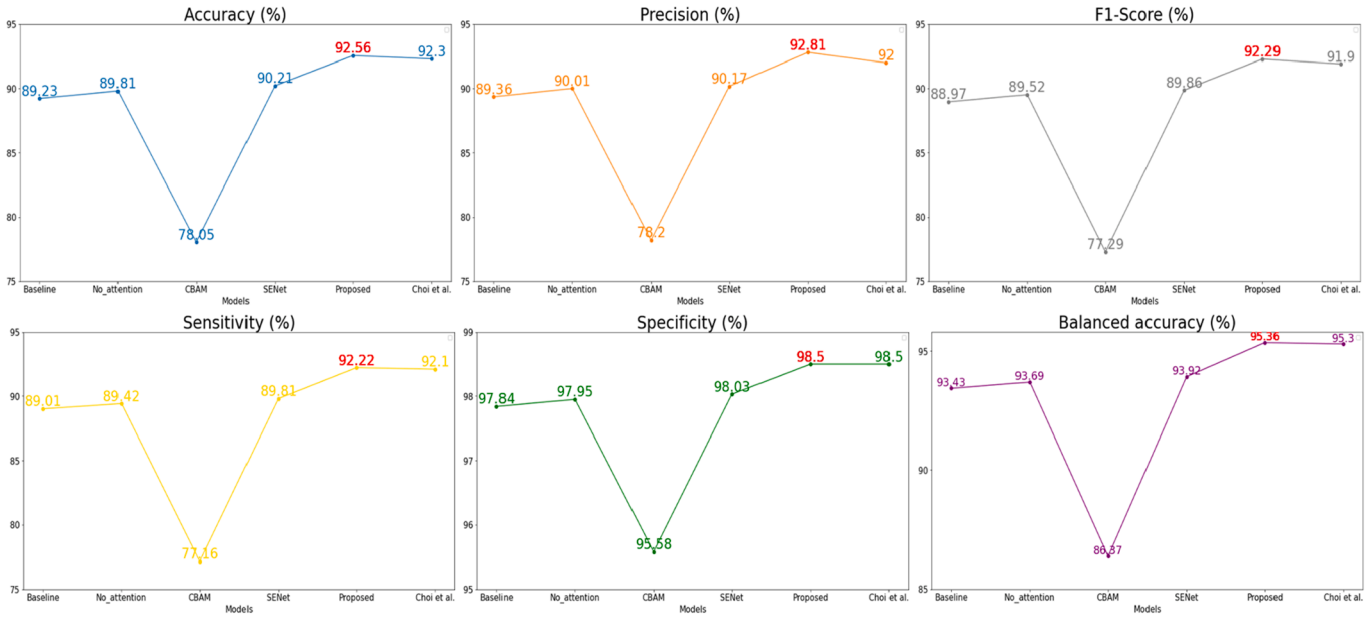


Fig. 8. Comparison of attention models: baseline, No-attention, CBAM, SENet, proposed model, and [23].

Table 11
Performance of pass filter: bandpass, lowpass, and highpass filters (%).

Filter	Accuracy	Precision	Sensitivity	Specificity	F1-score	Balanced accuracy
Bandpass filter (250 Hz-1800 Hz)	**92.56(±1.4)**	**92.81(±0.86)**	**92.22(±1.65)**	**98.5(±0.3)**	**92.29(±1.33)**	**95.36(±0.97)**
Lowpass filter (<250 Hz)	80.8(±1.58)	80.41(±1.55)	79.83(±1.6)	96.15(±0.32)	79.87(±1.57)	87.99(±0.96)
Highpass filter (>250 Hz)	91.28(±3.08)	91.67(±2.97)	90.91(±3.18)	98.24(±0.62)	91.07(±3.1)	94.57(±1.9)

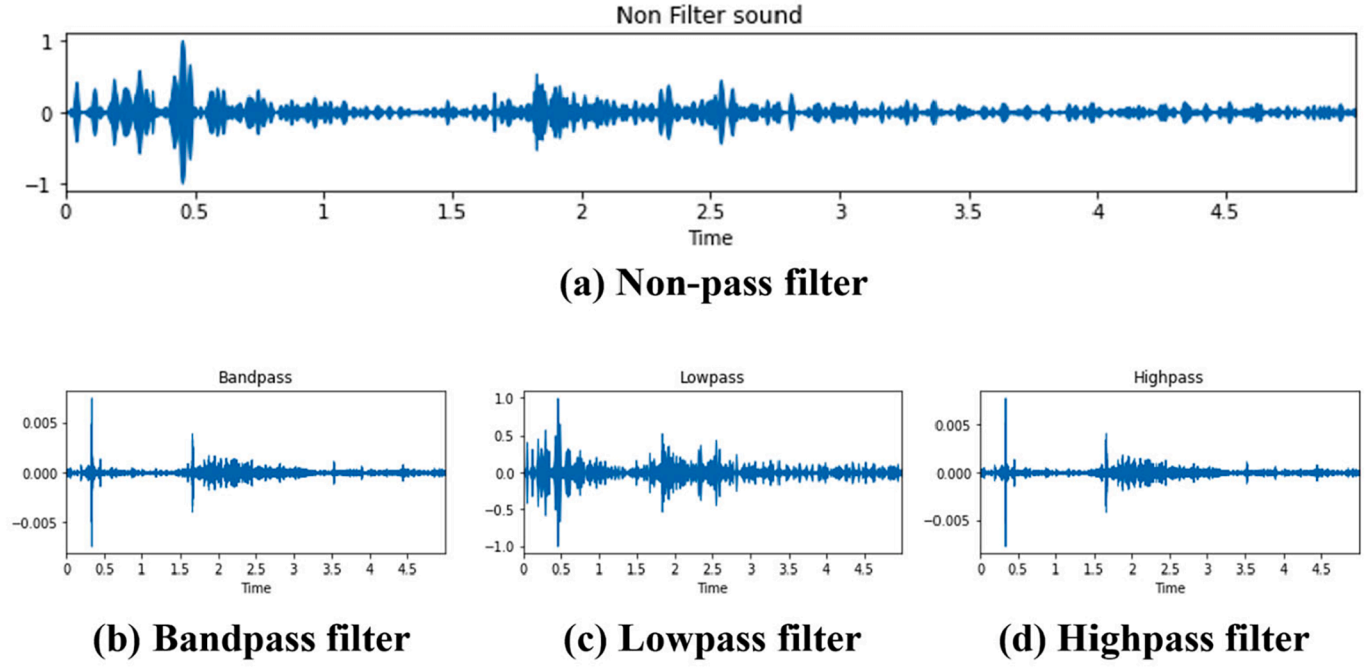


Fig. 9. Normal sound with pass filters: (a) non-pass filter, (b) bandpass filter, (c) lowpass filter, and (d) highpass filter.

LACM optimized for audio embedding. The main features were extracted using the log-Mel spectrogram MFCC of respiratory sounds collected using a smart stethoscope (Littmann 3200). Log-Mel spectrogram MFCC detects low and high-frequency bands of lung sounds—one of the advantages of the log-Mel spectrogram—and is suitable for learning data of respiratory sounds. MFCC also has the effect of removing noise through energy compression by DCT conversion. Combining the two methods, the main features of the appropriate band were extracted from the respiratory sound data collected in this study. We examined the visualization results of Grad-CAM to explain the evaluation of the proposed

Table 12
Performance of augmentation: time stretch and pitch (%).

Augmentation	Accuracy	Precision	Sensitivity	Specificity	F1-score	Balanced accuracy
Proposed model	**92.56(±1.4)**	**92.81(±0.86)**	**92.22(±1.65)**	**98.5(±0.3)**	**92.29(±1.33)**	**95.36(±0.97)**
w/ Time stretch (0.7)	88.93(±3.52)	89.51(±3.17)	88.39(±3.7)	97.76(±0.71)	88.7(±3.55)	93.07(±2.20)
w/ Time stretch (1.3)	89.81(±1.75)	89.96(±1.91)	89.68(±1.46)	97.95(±0.34)	89.69(±1.69)	93.81(±0.9)
w/ pitch	88.64(±4.09)	88.87(±4.28)	88.38(±4.15)	97.72(±0.8)	88.46(±4.28)	93.05(±2.47)

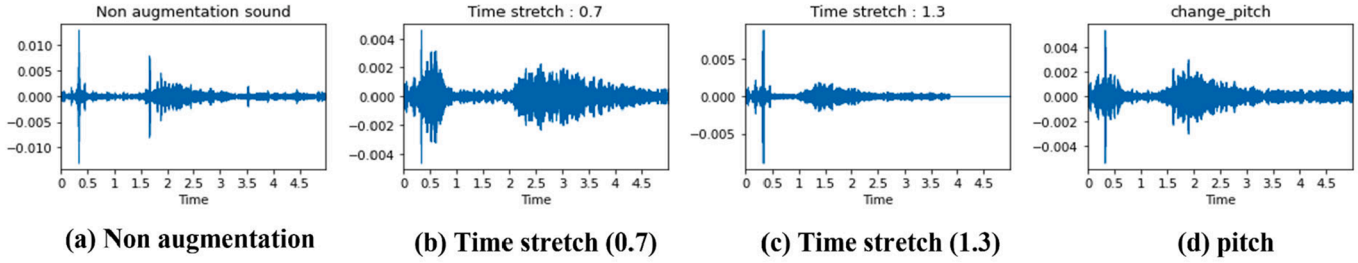


Fig. 10. Data augmentation: (a) non-augmentation, (b) time stretch (0.7), (c) time stretch (1.3), and (d) pitch.

Table 13
Lung sound of data [50].

Disease	Value
Normal	339
COPD	87
Asthma	300
Pneumonia	57
Bronchitis	24
Total	**807**

Table 14
Comparison with other models using the public dataset.

Models	Dataset	Significance	Accuracy (%)
Park et al. [51]	Littmann 3200 stethoscope	Mahalanobis / ResNet	80.57
Tripathy et al. [52]		EWT / LGBM	84.76
Soni et al. [53]		Contrastive learning external validation set	69.1(AUC)
Proposed Model		**Depthwise separable / LACM Attention**	**86.37(±SD 4.02)**

classification model. The classification model intensively learns feature information using the improved VGGish and attention modules. The classification model transformed the CNN structure of VGGish into a depthwise separable convolution to learn the voice features of the log-Mel spectrogram MFCC.

As the primary architecture of the model, the attention module is proposed to emphasize the characteristics of respiratory sounds. The attention module combines depthwise separable convolution and ECA-Net. The attention module takes advantage of ECA-Net's lightweight design with lower complexity than SENet and CBAM. We tried to make the model lighter overall. As shown in Table 15, the performance of the proposed classification model obtained an accuracy of 92.56 %, precision of 92.81 %, sensitivity of 92.22 %, specificity of 98.50 %, f1-score of 92.29 %, and balanced accuracy of 95.4 %. Compared to the previous study by Choi et al. [23], 6 class improved by 0.26 %, 0.81 %, 0.12 %, and 0.39 % in accuracy, precision, sensitivity, and f1-score, respectively. In addition, we included an experiment for 3 class and confirmed a performance improvement of 0.1 % and 0.3 % in precision and f1-score, respectively. We compared VGGish [39] and L3-Net optimized for audio embedding. The proposed model was obtained by modifying the VGGish structure, and an improvement in performance was demonstrated. When training two models, we changed the flattened layer to global average pooling. VGGish showed higher performance than L3-Net and was used in this study. Upon comparing the flops, the proposed model was found to be lighter.

The superiority of the model was confirmed through various comparative experiments, such as experimental comparison between attention modules, preprocessing of the required band, and application of sound augmentation. We confirmed that the combination of bandpass filter preprocessing, no augmentation, and LACM showed the highest performance. In addition, an accuracy of 86.37 % or more was achieved using the public dataset.

As for the limit of model performance, normal sound data had lower

Table 15
State-of-the-art models vis-a-vis the proposed model (%).

Related works	Models	Dataset		Accuracy	Precision	Sensitivity	Specificity	F1-score	Flops (billion)
		Dataset	**class**						
Park et al. [51]	Mahalanobis / ResNet	Littmann 3200 (Public dataset)	4	80.57	–	–	–	–	–
Tripathy et al. [52]	EWT / LGBM		2	84.76	–	–	–	–	–
Proposed model	Depthwise separable / LACM Attention		5	86.37	88.27	83.3	95.74	85.20	**1.0**
Hershey et al. [39]	VGGish	Clinical dataset	6	76.89	77.48	75.77	95.35	75.83	2.9
Arandjelovic et al. [20]	L3-Net			76.69	76.77	75.81	95.32	75.89	2.5
Choi et al. [23]	CNN BiGRU			92.3	92.0	92.1	98.5	91.9	–
			3	94.6	93.3	91.8	96.8	92.5	–
Proposed model	Depthwise separable / LACM Attention		**6**	**92.56**	**92.81**	**92.22**	**98.5**	**92.29**	**1.0**
			3	94.6	93.4	92.4	96.8	92.8	

performance than other classes. At that time, there was a limit to the collection of respiratory sound data of normal people because of the COVID-19 outbreak. Because we were targeting patients with lung disease, the normal sounds were relatively insufficient. However, the model's effectiveness was demonstrated by improving the classification performance of the low normal in the baseline and the performance improvement in the proposed model. Although there was a limit to securing patient data in the COVID-19 non-face-to-face situation, it presented an objective performance in lung disease classification.

The high performance of the classification model guarantees the superiority of algorithm performance. However, AI models are problematic because they do not interpret the black box in predictive classification. In many cases, it was impossible to explain how the classification performance was reached. Despite the high performance of classification prediction, visualization of learned features and accurate detection of medical patterns within the black box were difficult [54]. AI judgment errors in the medical field are fatal to patient health [55]. Therefore, we focused on the classification performance of models for medical decision-making and supported and interpreted XAI to make it more clinician-trustworthy. Grad-CAM was used to show the contribution of identifying feature points and patterns in respiration. This study used Grad-CAM to explain how to classify lung diseases with high performance. We used Grad-CAM to visualize normal sounds and patterns of five lung diseases, subsequently characterize the sections where abnormal sounds appear, and perform auxiliary analysis. In particular, the results mentioned in this study confirmed the phenomenon of cracking sounds and abnormal respiration for each of the five lung diseases. We received confirmation from the respiratory clinician regarding the experimental results and clinical opinions.

The advantage of this study is that in the respiratory sound sample, XAI well detected the region that can be considered as the interval of disease judgment. Respiratory sounds confirmed that normal sounds maintain constant breathing. Bronchiectasis, pneumonia, and ILD symptoms detected respiratory abnormalities. COPD and asthma detected abnormal respiratory sounds (wheezing) in the high band during exhalation because of airway stenosis. Therefore, this study showed the advantage of being able to be used as an aid to determine inhalation, exhalation, abnormal respiratory sounds found in visual analysis, and disease prediction. However, there was still a limit to using it as a complete decision-making process. Depending on the patient's condition, the lung field that can be diagnosed is different, and it is necessary to reflect some of the patient's clinical information rather than XAI problems. Improvements are needed based on additional data collection and the know-how and experience of medical staff.

There were advantages and limitations for XAI proposed in this study. Therefore, we plan a multi-modal concept prediction model based on clinical information by synthesizing clinical opinions. In addition, the correlation between each symptom and lung lesion can be expected using clinical information indicating the location of the lung lesion. Through this, it is expected that it will be possible to develop medical software that serves as a guide for medical staff who can judge diseases by converging the opinions of clinicians and XAI models.

7. Conclusion

Medical staffs have difficulty identifying abnormal respiratory sounds by auscultating sounds of lung disease and finding the characteristics or patterns of adventitious sounds. This is because respiratory sounds have a complex structure, so there is a limit to identifying patterns of various nonlinear data [14]. We proposed a lung disease classification model for patients suffering from lung disease. The proposed AI lung disease classification model shows more than 90 % of accuracy and is expected to play a significant role as an auxiliary tool to help medical staff decide whether there is an abnormality in breathing during inhalation or expiration pattern. It is valuable to use CNN models based on various measurement environments and experiences for each medical staff to detect diseases early and improve the patient's prognosis [19].

The dataset used in this study is meaningful in that two medical staff directly listened to respiratory sounds and provided labeled data for lung diseases. Respiratory sounds were measured from four lung lesions directly using a stethoscope, and patients with actual diseases were targeted. COPD and asthma were confirmed diseases in both lung fields, and pneumonia, bronchiectasis, and ILD were confirmed at the site. However, there is a limit to the diversity of medical devices. Data vary depending on the noise technology, stethoscope frequency, measurement location, medical staff's experience, and sound patterns. Collecting data using various stethoscopes and images is necessary to improve the model.

The proposed model is innovative and provides high accuracy and interpretability. The XAI described in this study is expected to contribute to the diagnosis of lung diseases by medical staff through early auscultation. In the future, we plan to develop lung disease-related research at the level of face-to-face treatment by fusing X-rays images and disease information. It is expected that cooperation between AI model experts and respiratory system specialists will bridge the medical gap and reduce the burden of high medical expenses.

8. Informed consent statement

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Clinical Research Ethics Committee of the Catholic Medical Center under Application No. KC20ONSI0774.

CRediT authorship contribution statement

Youngjin Choi: Conceptualization, Investigation, Methodology, Writing – original draft. **Hongchul Lee:** Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the Korea University Grant [No. k2209271, 2022] and in part by Brain Korea 21 FOUR. This paper is an extended version of "Lightweight Skip Connections With Efficient Feature Stacking for Respiratory Sound Classification" by Choi et al, published in IEEE Access.

References

[1] World Health Organization, World Health Statistics 2021: Monitoring health for the SDGs, sustainable development goals, Ind. High. Educ. 3 (2021) 1689–1699.
[2] G. Petmezas, G.-A. Cheimariotis, L. Stefanopoulos, B. Rocha, R.P. Paiva, A. K. Katsaggelos, N. Maglaveras, Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function, Sensors 22 (2022) 1232, https://doi.org/10.3390/s22031232.
[3] J.M. Leung, M. Niikura, C.W.T. Yang, D.D. Sin, Covid-19 and COPD, Eur. Respir. J. 56 (2020), https://doi.org/10.1183/13993003.02108-2020.
[4] Y. Ma, X. Xu, Y. Li, LungRN+ NL: an improved adventitious lung sound classification using non-local block ResNet neural network with mixup data augmentation, Interspeech (2020) 2902–2906, https://doi.org/10.21437/interspeech.2020-2487.
[5] B.M. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, A respiratory sound database for

the development of automated classification, Int. Conf. Biomed. Heal. Informatics, Springer (2017) 33–37, https://doi.org/10.1007/978-981-10-7419-6_6.
[6] T. Xia, J. Han, C. Mascolo, Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues, Exp. Biol. Med. (2022) 15353702221115428. https://doi.org/10.1177/15353702221115428.
[7] M. Sarkar, I. Madabhavi, N. Niranjan, M. Dogra, Auscultation of the respiratory system, Ann. Thorac. Med. 10 (2015) 158, https://doi.org/10.4103/1817-1737.160831.
[8] H. Pham Thi Viet, H. Nguyen Thi Ngoc, V. Tran Anh, H. Hoang Quang, Classification of lung sounds using scalogram representation of sound segments and convolutional neural network, J. Med. Eng. Technol. 46 (2022) 270–279. https://doi.org/10.1080/03091902.2022.2040624.
[9] R. Zulfiqar, F. Majeed, R. Irfan, H.T. Rauf, E. Benkhelifa, A.N. Belkacem, Abnormal respiratory sounds classification using deep CNN through artificial noise addition, Front. Med. 8 (2021), https://doi.org/10.3389/fmed.2021.714811.
[10] L. Shi, K. Du, C. Zhang, H. Ma, W. Yan, Lung sound recognition algorithm based on vGGish-BiGru, IEEE Access. 7 (2019) 139438–139449, https://doi.org/10.1109/access.2019.2943492.
[11] M. Fraiwan, L. Fraiwan, M. Alkhodari, O. Hassanin, Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory, J. Ambient Intell. Humaniz. Comput. 13 (2021) 4759–4771, https://doi.org/10.1007/s12652-021-03184-y.
[12] E. Grooby, C. Sitaula, D. Fattahi, R. Sameni, K. Tan, L. Zhou, A. King, A. Ramanathan, A. Malhotra, G.A. Dumont, Real-time multi-level neonatal heart and lung sound quality assessment for telehealth applications, IEEE Access. 10 (2022) 10934–10948, https://doi.org/10.1109/access.2022.3144355.
[13] J.A. Dar, K.K. Srivastava, S.A. Lone, Spectral features and optimal hierarchical attention networks for pulmonary abnormality detection from the respiratory sound signals, Biomed. Signal Process. Control. 78 (2022), 103905, https://doi.org/10.1016/j.bspc.2022.103905.
[14] M. Aykanat, Ö. Kılıç, B. Kurt, S. Saryal, Classification of lung sounds using convolutional neural networks, EURASIP J. Image Video Proc. 2017 (2017) 65, https://doi.org/10.1186/s13640-017-0213-2.
[15] O. Stephen, M. Sain, U.J. Maduh, D.-U. Jeong, An efficient deep learning approach to pneumonia classification in healthcare, J. Health. Eng. 2019 (2019), https://doi.org/10.1155/2019/4180949.
[16] K.J. Park, Y.J. Choi, H.C. Lee, COVID-19 CXR classification: applying domain extension transfer learning and deep learning, Appl. Sci. 12 (2022), https://doi.org/10.3390/app122110715.
[17] Y. Cao, C. Zhang, C. Peng, G. Zhang, Y. Sun, X. Jiang, Z. Wang, D. Zhang, L. Wang, J. Liu, A convolutional neural network-based COVID-19 detection method using chest CT images, Ann. Transl. Med. 10 (2022), https://doi.org/10.21037/atm-22-534.
[18] Y. Kim, Y. Hyon, S. Lee, S.-D. Woo, T. Ha, C. Chung, The coming era of a new auscultation system for analyzing respiratory sounds, BMC Pulm. Med. 22 (2022) 119, https://doi.org/10.1186/s12890-022-01896-1.
[19] Y. Kim, Y. Hyon, S.S. Jung, S. Lee, G. Yoo, C. Chung, T. Ha, Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning, Sci. Rep. 11 (2021) 17186, https://doi.org/10.1038/s41598-021-96724-7.
[20] R. Arandjelovic, A. Zisserman, Look, listen and learn, in Proc. IEEE Int. Conf. Comput. Vis., 2017: pp. 609–617. https://doi.org/10.48550/arXiv.1705.08168.
[21] A. Ponomarchuk, I. Burenko, E. Malkin, I. Nazarov, V. Kokh, M. Avetisian, L. Zhukov, Project Achoo: a practical model and application for COVID-19 detection from recordings of breath, voice, and cough, IEEE J. Sel. Top. Signal Process. 16 (2022) 175–187, https://doi.org/10.1109/jstsp.2022.3142514.
[22] G. Altan, Y. Kutlu, N. Allahverdi, Deep learning on computerized analysis of chronic obstructive pulmonary disease, IEEE J. Biomed. Heal. Inform. 24 (2020) 1344–1350, https://doi.org/10.1109/JBHI.2019.2931395.
[23] Y. Choi, H. Choi, H. Lee, S. Lee, H. Lee, Lightweight skip connections with efficient feature stacking for respiratory sound classification, IEEE Access (2022), https://doi.org/10.1109/access.2022.3174678.
[24] G. Altan, A. Yayık, Y. Kutlu, Deep learning with ConvNet predicts imagery tasks through EEG, Neural Process. Lett. 53 (2021) 2917–2932, https://doi.org/10.1007/s11063-021-10533-7.
[25] S. Gupta, M. Agrawal, D. Deepak, Gammatonegram based triple classification of lung sounds using deep convolutional neural network with transfer learning, Biomed. Signal Process. Control. 70 (2021), 102947, https://doi.org/10.1016/j.bspc.2021.102947.
[26] N. Asatani, T. Kamiya, S. Mabu, S. Kido, Classification of respiratory sounds using improved convolutional recurrent neural network, Comput. Electr. Eng. 94 (2021), 107367, https://doi.org/10.1016/j.compeleceng.2021.107367.
[27] S.B. Shuvo, S.N. Ali, S.I. Swapnil, T. Hasan, 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. Heal. Inform. (2020), https://doi.org/10.1109/jbhi.2020.3048006.
[28] L. Kranthi Kumar, P.J.A. Alphonse, COVID-19 disease diagnosis with light-weight CNN using modified MFCC and enhanced GFCC from human respiratory sounds, Eur. Phys. J. Spec. Top. (2022) 1–18, https://doi.org/10.1140/epjs/s11734-022-00432-w.
[29] G. Altan, Y. Kutlu, A.Ö. Pekmezci, S. Nural, Deep learning with 3D-second order difference plot on respiratory sounds, Biomed. Signal Process. Control. 45 (2018) 58–69, https://doi.org/10.1016/j.bspc.2018.05.014.
[30] G. Altan, Y. Kutlu, A. Gökçen, Chronic obstructive pulmonary disease severity analysis using deep learning on multi-channel lung sounds, Turkish J. Electr. Eng. Comput. Sci. 28 (2020) 2979–2996, https://doi.org/10.3906/ELK-2004-68.
[31] A. Qayyum, I. Razzak, M. Tanveer, A. Kumar, Depth-wise dense neural network for automatic COVID 19 infection detection and diagnosis, Ann. Oper. Res. (2021) 1–21, https://doi.org/10.1007/s10479-021-04154-5.
[32] Y. Chen, W. Du, X. Duan, Y. Ma, H. Zhang, Squeeze-and-excitation convolutional neural network for classification of malignant and benign lung nodules, J. Adv. Inf. Technol. 12 (2021), https://doi.org/10.12720/jait.12.2.153-158.
[33] A. Haghanifar, M.M. Majdabadi, Y. Choi, S. Deivalakshmi, S. Ko, Covid-cxnet: detecting COVID-19 in frontal chest x-ray images using deep learning, Multimed. Tools Appl. (2022) 1–31, https://doi.org/10.1007/s11042-022-12156-z.
[34] T. Zhang, G. Feng, J. Liang, T. An, Acoustic scene classification based on Mel spectrogram decomposition and model merging, Appl. Acoust. 182 (2021), 108258, https://doi.org/10.1016/j.apacoust.2021.108258.
[35] M.K. Das, S. Ari, Analysis of ECG signal denoising method based on S-transform, IRBM 34 (2013) 362–370, https://doi.org/10.1016/j.irbm.2013.07.012.
[36] Z. Tariq, S.K. Shah, Y. Lee, Feature-based fusion using CNN for lung and heart sound classification, Sensors 22 (2022) 1521, https://doi.org/10.3390/s22041521.
[37] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018: pp. 7132–7141. https://doi.org/10.1109/cvpr.2018.00745.
[38] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (2020) 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155.
[39] S. Hershey, S. Chaudhuri, D.P.W. Ellis, J. F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, CNN architectures for large-scale audio classification, in 2017 IEEE Int. Conf. Acoust. Speech Signal Process., IEEE, 2017: pp. 131–135. https://doi.org/10.1109/icassp.2017.7952132.
[40] E. Tsalera, A. Papadakis, M. Samarakou, Comparison of pre-trained CNNs for audio classification using transfer learning, J. Sens. Actuator Networks 10 (2021) 72, https://doi.org/10.3390/jsan10040072.
[41] Z. S. Syed, S.A. Memon, A.L. Memon, Deep acoustic embeddings for identifying Parkinsonian speech, Int. J. Adv. Comput. Sci. Appl. 11 (2020). https://doi.org/10.14569/ijacsa.2020.0111089.
[42] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Searching for MobileNetV3, in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019: pp. 1314–1324. https://doi.org/10.1109/iccv.2019.00140.
[43] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016: pp. 2921–2929. https://doi.org/10.1109/cvpr.2016.319.
[44] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in Proc. IEEE Int. Conf. Comput. Vis., 2017: pp. 618–626. https://doi.org/10.48550/arXiv.1610.02391.
[45] A.J. Larner, Accuracy of cognitive screening instruments reconsidered: overall, balanced or unbiased accuracy? Neurodegener. Dis. Manag. 12 (2022) 67–76, https://doi.org/10.2217/nmt-2021-0049.
[46] D. Choe, E. Choi, D.K. Kim, The real-time mobile application for classifying of endangered parrot species using the CNN models based on transfer learning, Mob. Inf. Syst. 2020 (2020), https://doi.org/10.1155/2020/1475164.
[47] G. Altan, Y. Kutlu, Y. Garbi, A.Ö. Pekmezci, S. Nural, Multimedia respiratory database (RespiratoryDatabase@ TR): auscultation sounds and chest X-rays, Nat. Eng. Sci. 2 (2017) 59–72, https://doi.org/10.28978/nesciences.349282.
[48] A. Bohadana, G. Izbicki, S.S. Kraman, Fundamentals of lung auscultation, N. Engl. J. Med. 370 (2014) 744–751, https://doi.org/10.1056/nejmra1302901.
[49] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, CBAM: Convolutional block attention module, in Proc. Eur. Conf. Comput. Vis., 2018: pp. 3–19. https://doi.org/10.1007/978-3-030-01234-2_1.
[50] M. Fraiwan, L. Fraiwan, B. Khassawneh, A. Ibnian, A dataset of lung sounds recorded from the chest wall using an electronic stethoscope, Data Br. 35 (2021), 106913, https://doi.org/10.1016/j.dib.2021.106913.
[51] C. Park, A. Awadalla, T. Kohno, S. Patel, Reliable and trustworthy machine learning for health using dataset shift detection, Adv. Neural Inf. Process. Syst. 34 (2021) 3043–3056, https://doi.org/10.48550/arXiv.2110.14019.
[52] R.K. Tripathy, S. Dash, A. Rath, G. Panda, R.B. Pachori, Automated detection of pulmonary diseases from lung sound signals using fixed-boundary-based empirical wavelet transform, IEEE Sensors Lett. 6 (2022) 1–4, https://doi.org/10.1109/lsens.2022.3167121.
[53] P.N. Soni, S. Shi, P.R. Sriram, A.Y. Ng, P. Rajpurkar, Contrastive learning of heart and lung sounds for label-efficient diagnosis, Patterns 3 (2022), 100400, https://doi.org/10.1016/j.patter.2021.100400.
[54] G. Altan, Deep OCT: an explainable deep learning architecture to analyze macular edema on OCT images, Eng. Sci. Technol. an Int. J. 34 (2022), 101091, https://doi.org/10.1016/j.jestch.2021.101091.
[55] J.K. Kim, S. Jung, J. Park, S.W. Han, Arrhythmia detection model using modified DenseNet for comprehensible Grad-CAM visualization, Biomed. Signal Process. Control. 73 (2022), 103408, https://doi.org/10.1016/j.bspc.2021.103408.