Attention Feature Fusion Network via Knowledge Propagation for Automated Respiratory Sound Classification

Ida A. P. A. Crisdayanti, Sung Woo Nam, Seong Kwan Jung, and Seong-Eun Kim, *Senior Member, IEEE*

Abstract*—*Goal: **In light of the COVID-19 pandemic, the early diagnosis of respiratory diseases has become increasingly crucial. Traditional diagnostic methods such as computed tomography (CT) and magnetic resonance imaging (MRI), while accurate, often face accessibility challenges. Lung auscultation, a simpler alternative, is subjective and highly dependent on the clinician's expertise. The pandemic has further exacerbated these challenges by restricting face-to-face consultations. This study aims to overcome these limitations by developing an automated respiratory sound classification system using deep learning, facilitating remote and accurate diagnoses.** ***Methods:*** **We developed a deep convolutional neural network (CNN) model that utilizes spectrographic representations of respiratory sounds within an image classification framework. Our model is enhanced with attention feature fusion of low-to-high-level information based on a knowledge propagation mechanism to increase classification effectiveness. This novel approach was evaluated using the ICBHI benchmark dataset and a larger, self-collected Pediatric dataset comprising outpatient children aged 1 to 6 years.** ***Results:*** **The proposed CNN model with knowledge propagation demonstrated superior performance compared to existing state-of-the-art models. Specifically, our model showed higher sensitivity in detecting abnormalities in the Pediatric dataset, indicating its potential for improving the accuracy of respiratory disease diagnosis.** ***Conclusions:*** **The integration of a knowledge propagation mechanism into a CNN model marks a significant advancement in the field of automated diagnosis of respiratory disease. This study paves the way for more accessible and precise healthcare solutions, which is especially crucial in pandemic scenarios.**

***Index Terms*—Classification, deep learning, feature fusion, knowledge propagation, respiratory sound.**

***Impact Statement*—Using a knowledge propagation mechanism, our deep learning model identifies respiratory abnormalities with high sensitivity and outperforms the accuracy of existing models, enabling effective remote diagnosis of respiratory diseases.**

I. Introduction

As INTERNET of Things (IoT) applications advance, sophisticated personal healthcare systems have emerged and continue to evolve. These systems are designed to enable early diagnosis of disease and continuous health monitoring even in the absence of trained medical staff. One key area of these systems is the automated classification of respiratory sounds for the early detection of abnormalities. Chronic respiratory disease (CRD), including chronic obstructive pulmonary disease (COPD), asthma, occupational lung diseases, and pulmonary hypertension, is the third leading cause of death globally, accounting for approximately 4 million deaths worldwide in 2019 [1]. While these diseases are not curable, various therapeutic interventions can significantly alleviate symptoms and enhance the quality of life for affected individuals [1]. Consequently, early detection of respiratory abnormalities is of paramount importance [2].

CRDs affect the airways and various structures within the lungs, resulting in abnormal breathing sounds [3]. Therefore, lung auscultation through respiratory sound analysis remains an indispensable method for the early detection of breathing irregularities. Although sophisticated diagnostic modalities such as X-ray imaging, ultrasonography, computed tomography (CT), and magnetic resonance imaging (MRI) are available, the stethoscope has been considered an invaluable tool for early-stage monitoring of CRDs [4], [5]. While auscultation is both non-invasive and cost-effective, it is limited by its inherent subjectivity and the specialized expertise required for the interpretation of abnormal respiratory sounds such as crackles and wheezing [6]. Furthermore, the recent COVID-19 pandemic has highlighted the importance of reliable telemedicine methods that allow remote diagnosis, thereby mitigating the risk of viral transmission through face-to-face consultations.

The inherent complexity associated with subtle differences between normal and abnormal respiratory sounds in a low-power spectrum presents challenges to clinicians. Given this complexity, machine learning (ML) and deep learning (DL) algorithms

Manuscript received 3 February 2024; revised 24 April 2024 and 13 May 2024; accepted 13 May 2024. Date of publication 16 May 2024; date of current version 7 June 2024. This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean Government under Grant RS-2023-00208492 and Grant 2020M3C1B8081320. The review of this article was arranged by Editor Adam C. Lammert. *(Corresponding author: Seong-Eun Kim.)*

Ida A. P. A. Crisdayanti and Seong-Eun Kim are with the Department of Applied Artificial Intelligence, Seoul National University of Science and Technology, Seoul 01811, South Korea (e-mail: sekim@seoultech.ac.kr).

Sung Woo Nam and Seong Kwan Jung are with the Woorisoa Children's Hospital, Seoul 08291, South Korea.

Digital Object Identifier 10.1109/OJEMB.2024.3402139

have increasingly attracted attention for their capacity to interpret intricate patterns in lung auscultation data recorded through advanced digital stethoscopes [7], [8], [9], [10], [11]. Automated respiratory sound analysis using ML and DL models can provide an early diagnosis of respiratory disease to individuals lacking specialized medical knowledge. Moreover, these automated systems offer more standardized measurements and consistent monitoring of preliminary symptoms, thereby expediting the initiation of preventative treatments [7], [12].

Various feature extraction techniques, such as Mel-Spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs), and scalograms, have been employed in ML and DL algorithms for respiratory sound detection [13], [14], [15], [16]. However, the efficacy of traditional ML-based methods depends substantially on the quality of the hand-crafted features. DL approaches, particularly convolutional neural networks (CNNs), offer a notable advantage through automated feature learning in the identification of anomalies from lung auscultation data [11], [13], [14], [17], [18]. The CNN models utilize spectrograms of respiratory sounds as input and demonstrate superior performance. This is attributed to hierarchical feature extraction occurring from the first convolutional layer through to deeper layers, evolving from low-level to higher-level feature representations [19]. These higher-level features represent global attributes of the spectrograms, aiding the classification layer in learning significant features from compressed information. However, because the process executes multiple convolutional filter operations, relevant information may be lost [20].

In this study, we developed a novel deep CNN model that utilizes feature fusion via knowledge propagation from high- to low-level features for respiratory sound classification. High-level features are extracted from the deepest convolutional layer to capture global representations, whereas low-level features encapsulate local information. The proposed hierarchical approach synergistically fuses high- and low-level features to improve classification performance. The process of knowledge propagation is performed iteratively by fusing a high-level feature with its preceding layer's local feature. This process culminates at the layer immediately above the stem layer, which is mainly responsible for the initial downsampling of input images to reduce computational cost [21]. Additionally, we introduce channel average pooling as a means to reduce dimensionality during feature fusion. In contrast with global average pooling, which computes the average across the time-frequency map, channel average pooling focuses only on the average of channels, thereby preserving time-frequency information and summarizing the information from multiple channels. Channel average pooling is enriched by a self-attention layer that reweights the extracted features to emphasize significant information. Employing a key-query paradigm, the self-attention mechanism calculates the attention scores for the respective weights of features [21]. This knowledge propagation technique can mitigate information loss by enriching global representations with significant localized features.

Our experimental results validate the efficacy of our knowledge propagation method in the form of enhanced performance of the deep CNN architecture for respiratory sound classification. Our approach established a new state-of-the-art performance on the International Conference on Biomedical Health Informatics (ICBHI) open dataset, achieving an improvement of approximately 1% in its score. Moreover, our model achieves a significantly increased accuracy in identifying abnormal samples, a crucial aspect in the detection of respiratory abnormalities. Furthermore, we rigorously evaluated our model on a self-collected dataset recorded from outpatient children aged 1 to 6 years, reaffirming its effectiveness.

In this study, we have made the following significant contributions:

1) We have developed a novel deep CNN model that incorporates knowledge propagation for feature fusion. This model combines high-level features (which capture global representations) with low-level features for localized information to accommodate the complexity inherent in the classification of respiratory sounds.
2) We have integrated channel average pooling as a dimensionality reduction technique (that preserves significant time-frequency information) with a self-attention layer that reweights features to emphasize significant information relevant to the classification.
3) We have validated the efficacy of our model using both the ICBHI open dataset and a self-collected Pediatric dataset of outpatient children aged 1 to 6 years. Through this rigorous validation process, we have established a new state-of-the-art performance.

The remainder of this paper is structured as follows: Section II contains a review of related work in the field of automated classification techniques. Section III contains a detailed discussion of our proposed method. In Section IV, experimental results validating the efficacy of our approach are presented. Finally, we conclude the paper in Section V.

II. Related Work

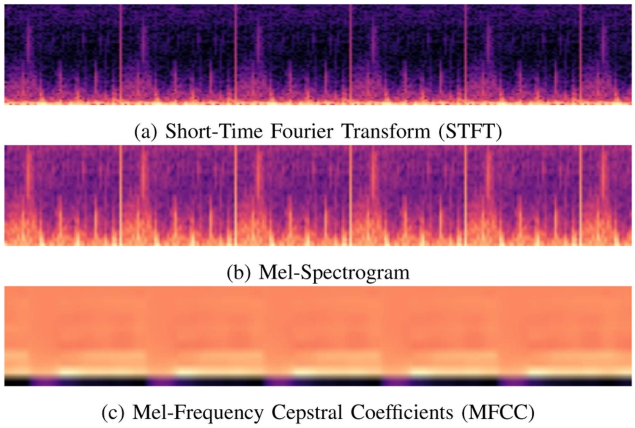
Adventitious sounds serve as key markers for abnormal respiratory conditions that overlap with normal breathing sounds [22]. Adventitious sounds are generally categorized as continuous (wheezes), discontinuous (crackles), or both [23], [24]. Wheezes are characterized by periodic and constant waveforms with a pitch above 100 Hz and a short duration [22], [25]. These sounds usually originate from airflow restrictions engaged by a narrow airway [3]. Wheezes are commonly associated with respiratory diseases such as asthma and COPD. Conversely, crackles indicate pathological abnormalities in the pulmonary tissue or airways [26]. These sounds exhibit explosive and discontinuous characteristics, known to occur during inspiration [22], [25]. Crackles are further subcategorized as fine or coarse, based on their duration [26]. Fine crackles, characterized by a short duration, are produced in the peripheral bronchi as a symptom of infection or pulmonary edema. Coarse crackles manifest at the beginning of an inhalation phase and indicate chronic bronchial diseases.

Recent studies have focused on the development of automated respiratory sound classification models that effectively distinguish between normal and abnormal sounds. Deep neural networks (DNNs), known for their exceptional performance in various applications such as natural language processing, speech

recognition, and image and video processing [27], [28], [29], have been actively applied to the classification of respiratory sounds. Initially, Kochetov et al. proposed a recurrent neural network (RNN)-based architecture to capture the temporal dynamics of sound data. However, this approach does not always fully extract the intrinsic time-frequency information of respiratory sound [10]. Therefore, respiratory sounds are transformed into spectrogram images, and CNNs have been mainly used as the backbone architecture [11], [17]. Kim et al. leveraged the spatial locality properties of CNNs to classify the Mel-Spectrogram of respiratory sounds [11]. Further enhancements have been achieved through pre-training a deep CNN on large audio datasets [30]. The CNN6 model was pre-trained on the AudioSet dataset [31], which contains two million sounds, including respiratory sounds like breathing, coughing, and sneezing. This pre-training approach led to a performance enhancement of almost 3% compared to models without pre-training [30]. Messner et al. combined the respective strengths of RNNs and CNNs to exploit spectral, temporal, and spatial information from respiratory sounds [32]. This approach fed spectrograms into a convolutional recurrent neural network architecture to classify normal and abnormal respiratory sounds. Long short-term memory (LSTM) was then leveraged as the RNN module to improve feature memorization [33].

Recently, large CNN models pre-trained on the ImageNet dataset, such as ResNet, have demonstrated promising performance in respiratory sound classification. Gairola et al. proposed a respiratory sound classifier called RespireNet by utilizing ResNet34 as the backbone architecture and fine-tuning the final fully-connected layer on the respiratory sound dataset [17]. To enhance the effectiveness of transfer learning with limited task-specific data, data augmentation mechanisms were employed [17], [34]. In contrast with image data augmentation, standard sound data augmentation techniques, including noise addition, speed variation, random shifting, and pitch shifting, were applied to audio signals before transformation [17]. ResNeSt [35] was also fine-tuned to improve its classification performance, introducing circular padding as an augmentation technique to address the challenge of the imbalanced dataset by increasing the number of abnormal samples [34]. Rather than focusing on augmentation techniques, Nguyen et al. proposed a co-tuning method with a pre-trained ResNet50 model, enhancing the transfer learning process [36]. Co-tuning is a two-step framework where the first step involves learning the relationship between source and target categories from the pre-trained model with calibrated predictions. In the second step, target labels (one-hot labels) and source labels (probabilistic labels) collaboratively supervise the fine-tuning process as translated by the category relationship.

Although pre-trained models such as ResNet have demonstrated excellent performance in general image classification, they have predominantly been trained on datasets comprising images of objects such as animals, humans, and household items. The characteristics of these images differ from those of spectrograms transformed from audio signals. Spectrograms represent complex time-frequency relationships inherent in sound data, which may not be effectively captured by models pre-trained on general image datasets. Therefore, we propose a specific architecture based on CNNs designed to accommodate the distinct spectral patterns and temporal dynamics of respiratory sound data, thereby potentially providing improved classification performance compared to pre-trained models on general image datasets.



(a) Short-Time Fourier Transform (STFT)

(b) Mel-Spectrogram

(c) Mel-Frequency Cepstral Coefficients (MFCC)

Fig. 1. Audio signal conversion to image representation as input for deep learning models.

III. Methods

A. Sound-to-Image Transformation

To facilitate the classification of respiratory sounds using CNNs, the respiratory sound data were transformed into a particular feature representation suitable for image-based processing. We evaluated three transformation techniques: Short-Time Fourier Transform (STFT), Mel-Spectrogram, and Mel-Frequency Cepstral Coefficients (MFCC), each providing a unique visual representation of sound characteristics. The STFT captures dynamic time-frequency information, the Mel-Spectrogram emphasizes auditory perception by scaling frequencies according to the Mel scale (which approximates human hearing sensitivity), and the MFCC encapsulates the spectral envelope of sounds, which is essential for speech-related nuances. Fig. 1 illustrates image representations of these transformations as applied to a sample of respiratory audio recordings.

An audio signal is comprised of several single-frequency sound waves, and the frequency of the information varies over time. Consequently, the time-varying properties of the sound spectrum need to be represented. The Fourier transform, a fundamental tool in signal processing, is used to convert an audio signal from its time domain into a frequency domain representation, generating the signal's spectrum. To compute a spectrogram, which is a time-varying spectrum representation, an STFT is applied, where the audio signal is first sliced into N frames with overlapping windows, applying the Hamming window to minimize edge effects. The STFT computes the power spectrum, P_i, at each frequency bin i in accordance with

the following equation:

$$P_i = \frac{|S(i)|^2}{N}, \tag{1}$$

where $S(i) = \sum_{n=0}^{N-1} s(n) \exp(-j\frac{2\pi}{N}ni)$ are the Fourier coefficients of the windowed sound frame, $s[n]$. The resulting spectrogram visually represents the signal's loudness or amplitude, as it varies over time at different frequencies.

For the Mel-Spectrogram, the frequencies are converted to the Mel scale, better reflecting human auditory perception. The Mel-Spectrogram emphasizes finer resolution at lower frequencies and coarser resolution at higher frequencies with conversion formula $m = 2595 \log_{10}(1 + \frac{f}{700})$, where f is the frequency in Hertz and m is its Mel scale.

To obtain MFCCs, a discrete cosine transformation (DCT) is applied to the logarithmic Mel-spectrum, $S(m)$, extracting the cepstral coefficient and effectively separating the pitch information from the formants through the following formula:

$$Y(n) = \sum_{m=0}^{M-1} \log(S(m)) \cos\left(n\,(m + 0.5)\,\frac{\pi}{M}\right), \tag{2}$$

where M is the number of Mel frequency bands. This cepstral representation is especially appropriate for speech recognition applications because of its efficacy in capturing phonetically significant features while being less susceptible to noise.

For the implementation of the sound preprocessing, we followed the existing framework used in RespireNet [17], which involves transforming sounds into Mel-Spectrograms. This process uses a sampling rate of 4000 Hz, 64 Mel filterbanks, and an FFT window of 256 points. We also performed blank region clipping to remove unnecessary high-frequency regions, ensuring the network focuses on the most relevant information for improved performance [17]. In respiratory sound analysis, transforming the one-dimensional audio signal into a two-dimensional form is essential to take advantage of the advanced pattern recognition capabilities of CNNs. The key rationale for selecting an appropriate transformation method is to facilitate the detection of subtle variations in respiratory sounds, ensuring that the differences between normal and abnormal conditions, which might be challenging to discern in their original form, become more accessible and identifiable in the transformed two-dimensional representation.

B. Feature Fusion Through Knowledge Propagation

In the proposed deep CNN-based architecture, we implemented a feature fusion mechanism using our novel knowledge propagation concept to enhance classification performance. This mechanism strategically combines multi-level feature maps, enabling the network to maintain essential spectral-temporal information throughout layers, thereby enhancing the robustness and discriminative power of the learned features for respiratory sound classification.

The input to the model is the audio spectrogram, represented as $x \in \mathbb{R}^{T\times F\times C}$, where T and F denote the time and frequency dimensions, respectively. The number of channels, C, is equal to three to accommodate the RGB representation of the spectrogram image. As depicted in Fig. 2, the architecture consists of nine CNN blocks.

The three-channel RGB spectrogram image is fed into the first convolutional layer, termed the stem layer, and the output of the stem layer at each channel is given by:

$$f_{i,j} = \sum_{a=0}^{k-1}\sum_{b=0}^{k-1} w_{a,b} x_{i+a-1,j+b-1}, \tag{3}$$

where $k = 3$ is the kernel size, ws are learnable weights of the convolutional filter, and (i, j) is the index of the resulting feature map after the convolutional operation. Subsequently, batch normalization (BN) is performed and a rectified linear unit (ReLU) activation function is applied, expressed as:

$$\tilde{f}_{i,j} = ReLU(BN(f_{i,j})) = \max(0, BN(f_{i,j})). \tag{4}$$

The batch normalization incorporates the mean and standard deviation that are calculated per dimension over the mini-batches. The stem layer's primary role is spatial downsampling; its feature is comprised of RGB pixels that are individually uninformative and, therefore, exhibit high spatial correlation [21]. Because of these properties, the stem layer is not included in the knowledge propagation process. Max pooling with a 2×2 kernel and stride of 2 is applied at the end of the stem block B^{stem} for further spatial downsampling:

$$y^{\text{stem}}_{p,q} = MaxPooling(\tilde{f}_{p,q}) = \max_{a,b=0}^{1} \tilde{f}_{2p+a,2q+b}. \tag{5}$$

The stem block architecture is followed by two convolutional blocks with two different filter sizes. The earlier block employs a 1×1 convolutional filter with a gradually increasing number of output channels to capture various forms of information in the images. This channel-enhancement (CE) block was designed on the basis of studies that a large number of output channels generate higher performance [37]. The subsequent spatial-downsampling (SD) block is composed of a 3×3 convolutional filter that performs simultaneous feature extraction and spatial downsampling. The combined CE and SD blocks extract significant spatial features for classification. Both blocks incorporate batch normalization (for network regularization) and the ReLU activation function (which is robust against saturation and vanishing gradients, unlike the sigmoid and tanh functions [38]). The CE and SD blocks are stacked four times and the ith combined block is denoted as $B_i^{\text{CE-SD}}$.

The output of the stem block, y^{stem}, passes through the first CE and SD blocks sequentially, resulting in the extracted feature $y_1 = B_1^{\text{CE-SD}}(y^{\text{stem}})$. The feature y_1 is repeatedly fed into the next CE and SD blocks, and the global feature y_4, extracted from the final block, is represented as $y_4 = B_4^{\text{CE-SD}}(y_3)$. The global feature y_4 contains a compressed form of the input, distilling essential information into a compact representation that includes general knowledge of the sound input, thereby facilitating effective classification.

However, as the convolutional layers progress and reduce dimensions to extract higher-level features, there is an inherent risk of losing fine-grained details present in the initial layers. To

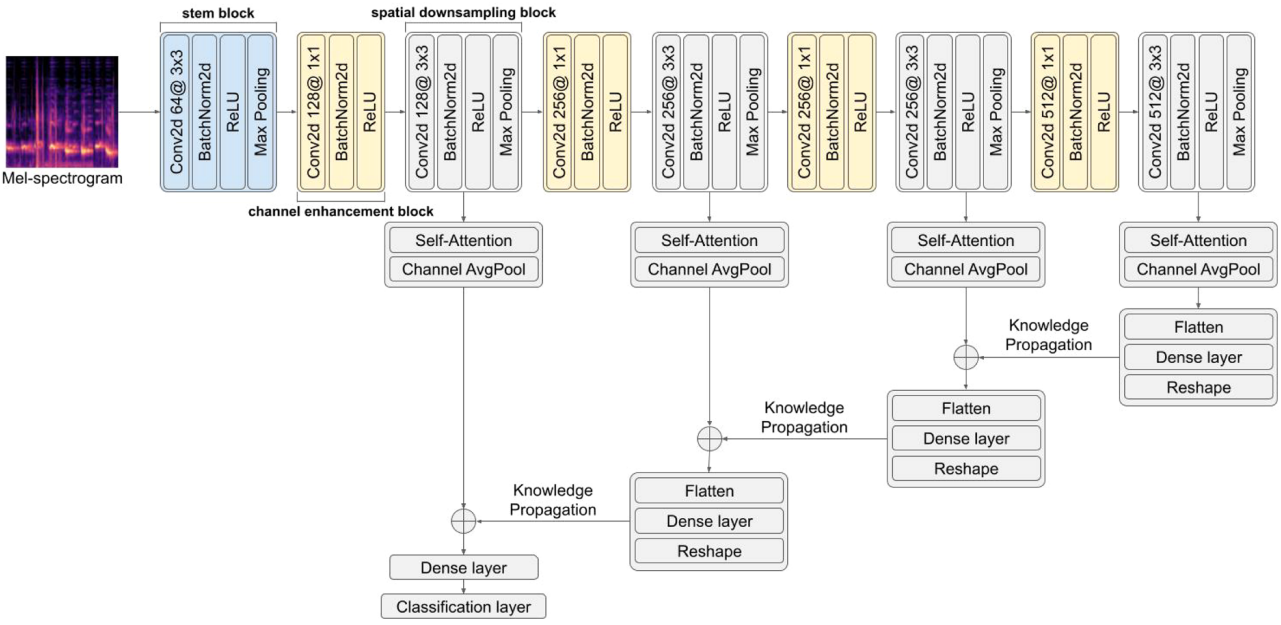


Fig. 2. The overall architecture of the proposed method.

address this possible information loss and ensure the robustness of features for sound classification, we developed a feature fusion mechanism. This mechanism was designed to integrate the global, high-level features with the detailed, lower-level features retained from earlier layers. The mechanism assists in preserving a comprehensive representation of input data and enhances the classifier's ability to distinguish between subtle variations in respiratory sounds.

As shown in Fig. 2, our model initiates knowledge propagation through feature fusion by integrating features from the fourth and third SD blocks. This integration of global and local features from successive levels of the network enables the propagation of informative features across the network. The fusion operation repeatedly combines features from the current and preceding SD blocks, thus preserving knowledge throughout the layers down to the first block.

Prior to the knowledge-propagation operation, the output of the combined CE and SD blocks is fed to a self-attention layer, $Attn$, to perform reweighting on relevant features, followed by a channel average pooling, CAP, expressed as follows:

$$\tilde{y}_n = CAP(Attn(B_n^{\text{CE-SD}}(y_{n-1}))). \tag{6}$$

Details of the self-attention and channel average pooling are elucidated in subsequent sections. Because of a series of convolutional filter operations, $\tilde{y}_n$ and $\tilde{y}_{n-1}$ have different dimensions. Therefore, a linear projection is applied to $\tilde{y}_n$ via a dense layer to align with the dimension of $\tilde{y}_{n-1}$ before the fusion. The 2D feature $\tilde{y}_n$ is first flattened from $\mathbb{R}^{1\times m\times n}$ to a 1D array in $\mathbb{R}^{1\times mn}$. Followed by a linear projection, the dimension is enlarged to $\mathbb{R}^{1\times pq}$ where p and q are the dimensions of the 2D feature $\tilde{y}_{n-1}$. The 1D feature output of the dense layer is subsequently reshaped to the 2D feature $\tilde{y}_n^{\text{up}} \in \mathbb{R}^{1\times p\times q}$ for the fusion operation. The fused features $\tilde{y}_{(n,n-1)}$ are obtained by:

$$\tilde{y}_{n,n-1} = \tilde{y}_n^{\text{up}} + \tilde{y}_{n-1}. \tag{7}$$

The final fused feature will contain both the global and the local features from different depths of convolutional blocks. This feature is then fed to a fully-connected layer for classification. To mitigate the risk of overfitting, dropout layers are applied following each dense layer in the model. The model is trained using the cross-entropy loss function, expressed as follows:

$$\text{loss} = -\sum_{c=1}^{C} l_i^c \log(p_i^c), \tag{8}$$

where C is the number of classes, l_i^c is a binary indicator (0 or 1) if class label c is the correct classification for sample i, and p_i^c is the predicted probability that sample i is of class c.

C. Self Attention

The self-attention mechanism is performed on the extracted features obtained from different convolutional layers. The self-attention layer is applied over a two-dimensional feature to perform reweighting on the input where the keys and values are the linear projection of the same features. This self-attention operation follows the formulation introduced by Ramachandran et al. [21]. Before the attention operation is performed over the two-dimensional features, a zero pad with a width of 3 is added along the input edges. After that, a 7×7 attention window with a single stride is applied to the two-dimensional features.

The center of the attended region is the query in the self-attention operation. For $x_{i,j}$ as the center of an attended region $a \times b$, the output of $Attn(x_{a,b})$ for an input channel can be

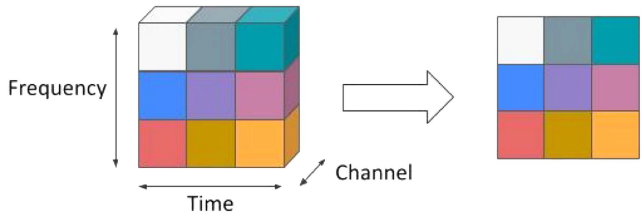


Fig. 3. Channel average pooling process.

obtained by the following formula:

$$Attn(x_{a,b})_{i,j} = \sum_{a,b \in N_k(i,j)} softmax_{a,b}(q_{i,j} \cdot k_{a,b}) \cdot v_{a,b} \quad (9)$$

where the query, $q_{i,j} = W_q x_{i,j}$, keys, $k_{a,b} = W_k x_{a,b}$, and values, $v_{a,b} = W_v x_{a,b}$, are linear transformations of the attended image features $a \times b$.

D. Channel Average Pooling

Before knowledge propagation, channel average pooling is an essential step in reducing feature dimensions while preserving significant information about the audio data. After the self-attention layer in our model produces reweighted 3D features, channel average pooling is applied to transform these 3D features into 2D features by summarizing channel information. In contrast with global average pooling, which averages over the entire time-frequency map, channel average pooling computes the average over the channel dimension, as illustrated in Fig. 3. This can be expressed as:

$$CAP(x)_{i,j} = \frac{1}{C} \sum_{k=1}^{C} x_{i,j,k} \quad (10)$$

where x represents the 3D input, (i, j) is the index of the output, and k denotes the channel index. This method effectively preserves the essential time-frequency information inherent in audio data with reduced feature dimensions.

Our model was trained for 200 epochs with a batch size of 16 and a constant learning rate of 0.001. Hyperparameters were optimized through a grid search approach. The best model weights for evaluation were chosen based on the highest performance on the validation set during the training phase.

IV. Results and Discussion

A. Dataset and Evaluation Metrics

To evaluate the performance of our proposed method, we trained and tested the model on a self-collected Pediatric dataset from Woorisoa Children's Hospital in South Korea and the publicly available ICBHI 2017 challenge dataset. The Pediatric dataset consists of pediatric respiratory sounds recorded by medical professionals from outpatient children aged 1 to 6 years at Woorisoa Children's Hospital. This recording was approved by the local institutional review board (Seoul National University of Science and Technology, No. 2021-0017). Written informed consent was obtained for the study. Two pediatricians carefully labeled a large amount of respiratory sound data as normal and abnormal. The characteristics of the hospital dataset are presented in Table 1. The ICBHI dataset contains 3,642 audio recordings classified as normal and 3,254 classified as abnormal, mostly recorded from adults.

TABLE 1
CHARACTERISTICS OF THE PEDIATRIC DATASET COLLECTED BY WOORISOA CHILDREN'S HOSPITAL

Variables	Value
Number of patients	675
Sex (Male/Female)	333/342
Age (Mean±SD)	3.18±1.48
Number of sounds (Normal/Abnormal)	5,605/5,549

For the Pediatric dataset, a 5-fold cross-validation method was used, ensuring that each fold has an equal distribution of normal and abnormal samples in each fold while keeping subjects between folds non-overlapping. This 5-fold cross-validation process was performed with the optimized hyperparameters and independently repeated five times to obtain a reliable performance evaluation. For the ICBHI dataset, we used the official split with a training set comprising 60% of the dataset and a test set comprising the remaining 40% for fair comparison [36]. This training-testing scheme was also repeated five times, and the average performance was obtained [30].

For a balanced performance evaluation, we evaluated the performance of the models using four evaluation metrics: accuracy, precision, recall, and F1-Score. In binary classification, accuracy is the ratio of correctly classified samples to the total number of samples. Precision is the ratio of correctly classified abnormal samples to the total samples predicted as abnormal. Recall is the ratio of correctly classified abnormal samples to the actual total of abnormal samples. The F1-score is the harmonic mean of precision and recall, providing a balance between them. To maintain comparability with baseline algorithms, we also presented sensitivity, specificity, and score. Specificity is the ratio of correctly classified normal samples to the total number of normal samples. Similarly, sensitivity is computed as the ratio of correctly classified abnormal samples to the total number of abnormal samples. As an overall measure of performance, the score is derived by averaging the two metrics, providing a balanced view of the classifier's capabilities.

To assess the statistical significance of our results, we conducted the Mann-Whitney U test. This non-parametric method is used to compare differences between two independent groups when the dependent variable is not normally distributed. As the test makes no assumptions about the distribution of the data, it is particularly well-suited for our study with small sample sizes.

B. Experimental Results

1) Sound Data Representation: In the respiratory sound classification, the choice of sound data representation is crucial to the overall performance of the classification model. To

TABLE 2
PERFORMANCE COMPARISON OF DIFFERENT INPUT REPRESENTATIONS ON ICBHI DATASET

Input	Acc	Precision	Recall	F1-Score
STFT	0.5297	0.4776	0.3338	0.3930
Mel-Spectrogram	**0.6597**	**0.6398**	**0.6118**	**0.6252**
MFCC	0.5314	0.4882	0.4730	0.4805

determine the most effective input representation for our proposed architecture, we conducted experiments using three different sound data representations: STFT, Mel-Spectrogram, and MFCC. As indicated in Table 2, the Mel-Spectrogram emerged as the superior representation, outperforming the other representations with a significant improvement of 23% in the F1-score. This result confirms the effectiveness of the Mel-Spectrogram in capturing essential features for the classification of respiratory sounds. Despite the fact that both the Mel-spectrogram and the MFCC utilized the Mel scale conversion, the MFCC demonstrated a comparatively lower performance. This may be attributed to the fundamental nature of MFCC, which focuses on emphasizing phonological information by discarding pitch information from the formants. Such a feature is suitable for speech processing tasks, but (as our results indicate) less effective for respiratory sound classification. The distinct acoustic properties of respiratory sounds, which may rely heavily on pitch information, probably contribute to the reduced effectiveness of MFCC in this context.

2) Performance Comparison on ICBHI Dataset: We evaluated our proposed model using the ICBHI open dataset by employing the Mel-Spectrogram as the input representation. To validate our model's superiority, we compared its performance with the current state-of-the-art models. For fair comparison, we re-evaluated the RespireNet models under the official 60:40 split dataset, using the open-source code provided by the authors [17]. The results, as listed in Table 3, demonstrate that our proposed method exhibits competitive performance, achieving an average score of 0.6569 across five repeated experiments. Specifically, the model displayed high sensitivity, a critical metric in medical diagnosing applications, as it reflects the model's efficacy in accurately identifying abnormal respiratory sounds. Given that the primary application of this study is to facilitate preliminary screening for respiratory diseases, a high rate of detection of abnormalities is paramount. With a 9% improvement in the sensitivity and a higher overall performance score, the proposed method achieved a record level of state-of-the-art performance on the normal-abnormal respiratory sound classification for the ICBHI dataset. As depicted in Fig. 4, the receiver operating characteristic (ROC) curve corroborates the superiority of our proposed method, compared to the RespireNet algorithms, by achieving a value of 0.70 for the area under the curve (AUC). Our model also shows remarkable enhancements in performance compared to other pre-trained models, such as VGG16 and DenseNet. The Mann-Whitney U test results between the proposed method and each baseline model reveal a p-value of 0.004 (< 0.05), which confirms that our proposed method significantly outperforms the baseline models.

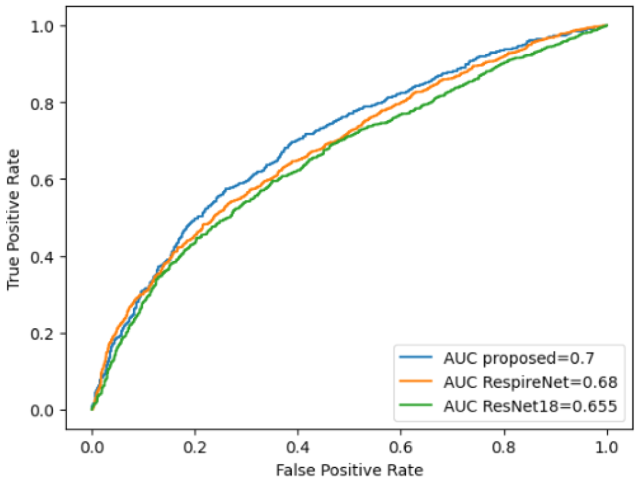


Fig. 4. ROC plot of RespireNet and the proposed method applied to ICBHI dataset.

3) Performance Comparison on Pediatric Dataset: In a further evaluation of our proposed deep learning model, we assessed its effectiveness using a self-collected Pediatric dataset; the results are listed in Table 4. The model consistently achieved an average sensitivity of 87.53%, confirming its robustness in accurately identifying abnormal respiratory sounds. In parallel with its high sensitivity, our model showed a higher overall performance score compared to the baseline RespireNet model. Notably, our model exhibited remarkably better performance on the Pediatric dataset compared to the ICBHI dataset, which includes data recorded from a wide age range across both old and young patients and collected at multiple hospitals using various recording devices. The diversity introduces a significant variance in the dataset, leading to more complex differentiation between normal and abnormal respiratory sounds. These variations present considerable challenges for the model's ability to generalize and accurately classify sounds. In contrast, the Pediatric dataset is more homogeneous, comprising data from children aged 1 to 6 years, all recorded using a single device. This homogeneity simplifies the classification task, resulting in better performance.

We trained the RespireNet using the Pediatric dataset with the 5-fold cross-validation, repeated five times. The results of each 5-fold cross-validation trial are presented in Table 4. These comprehensive tests reveal that our model consistently outperformed RespireNet, scoring 1.13% higher. The ROC curve, depicted in Fig. 5, further demonstrates that the proposed method effectively discriminates between normal and abnormal respiratory samples. In comparison to other pre-trained models, such as VGG16 and DenseNet, our model demonstrates a substantial superiority, achieving an F1-score improvement of over 14%. To assess the statistical significance of these performance differences, the Mann-Whitney U test was performed between the proposed method and each baseline model. This test yielded a p-value of 0.004 (< 0.05) across all comparisons, confirming the enhancement provided by our approach in respiratory sound classification.

TABLE 3
PERFORMANCE COMPARISON WITH THE BASELINE MODELS ON ICBHI DATASET

Methods	Accuracy	Specificity	Sensitivity	Score	Precision	F1-Score
VGG16	0.6154±0.007	0.6457±0.035	0.5826±0.042	0.6141±0.007	0.5871±0.010	0.5840±0.019
DenseNet	0.6353±0.014	0.6837±0.022	0.5757±0.026	0.6297±0.007	0.6113±0.009	0.5927±0.013
Vanilla Fine-tuning	0.6511	0.7633	0.5212	0.6422	0.6553	0.5806
StochNorm	0.6539	0.7886	0.4979	0.6432	0.6703	0.5714
CoTuning	0.6581	**0.7934**	0.5014	0.6474	**0.6769**	0.5761
CoTuning-StochNorm	0.6471	0.7856	0.4867	0.6361	0.6621	0.5610
RespireNet (ResNet34)	0.6587±0.005	0.7847±0.023	0.4919±0.029	0.6383±0.005	0.5206±0.016	0.5624±0.014
RespireNet (ResNet18)	0.6541±0.010	0.7556±0.014	0.5195±0.016	0.6375±0.010	0.6473±0.014	0.5763±0.013
Proposed Method	**0.6597±0.012**	0.7019±0.029	**0.6118±0.019**	**0.6569±0.011**	0.6398±0.018	**0.6252±0.011**

TABLE 4
5-FOLD CROSS-VALIDATION RESULTS ON THE PEDIATRIC DATASET

Methods	Accuracy	Specificity	Sensitivity	Score	Precision	F1-Score
VGG16	0.6918±0.008	0.6937±0.060	0.6862±0.073	0.6899±0.010	0.7704±0.057	0.7254±0.065
DenseNet	0.7207±0.016	0.8153±0.011	0.6154±0.014	0.7154±0.007	0.7498±0.010	0.6759±0.010
RespireNet (ResNet34)	0.8644±0.002	**0.9347±0.005**	0.7928±0.006	0.8638±0.002	**0.9161±0.005**	0.8500±0.003
RespireNet (ResNet18)	0.8591±0.003	0.9023±0.006	0.8130±0.005	0.8593±0.003	0.8942±0.002	0.8464±0.005
Proposed Method	**0.8755±0.007**	0.8748±0.021	**0.8753±0.010**	**0.8751±0.006**	0.8632±0.018	**0.8690±0.006**

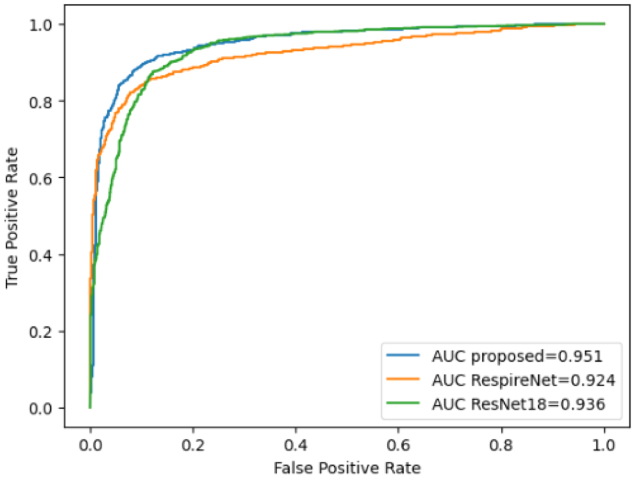


Fig. 5. ROC plot of RespireNet and the proposed method applied to the Pediatric dataset.

TABLE 5
ABLATION STUDY ON THE DEPTH OF KNOWLEDGE PROPAGATION. (A) ICBHI DATASET. (B) PEDIATRIC DATASET

(a)

Model	Acc	Precision	Recall	F1-Score
3 steps	**0.6597**	0.6398	**0.6118**	**0.6252**
2 steps	0.6517	**0.6448**	0.5244	0.5784
1 step	0.6305	0.6110	0.5475	0.5775
no propagation	0.6379	0.6339	0.4596	0.5329

(b)

Model	Acc	Precision	Recall	F1-Score
3 steps	**0.8755**	**0.8632**	**0.8753**	**0.8690**
2 steps	0.8609	0.8602	0.8449	0.8525
1 step	0.8500	0.8576	0.8221	0.8395
no propagation	0.8340	0.8084	0.8494	0.8284

C. Ablation Study

1) Effect of Propagation Steps: In order to gain a comprehensive understanding of the contributions of the various components in our proposed model, we conducted an ablation study. This involved removing certain modules or operations from the overall architecture and observing the resultant effects on performance. The original architecture employed a three-step knowledge propagation, where high-level information from the deepest CNN layer was propagated to the preceding layers. The model with 1-step knowledge propagation refers to propagation occurring between the last two spatial downsampling blocks. 2-step and 3-step knowledge propagation extend this process by incorporating features from additional preceding spatial downsampling blocks, following the backward direction illustrated in Fig. 2. As described in Table 5, when testing our model on the ICBHI dataset, reducing the number of knowledge propagation steps decreased the accuracy, precision, recall, and F1-Score in general. Consequently, the no-propagation model, which classifies features obtained from the last CNN layer, achieved the lowest F1-score.

These results suggest that the depth of knowledge propagation has a direct impact on the richness of the fused knowledge. The integration of local features with global information becomes more effective with deeper propagation. High-level features, which encapsulate global information, are robust against noise but may lack essential details due to oversimplification [20]. In contrast, low-level features provide local, rich, detailed information but are more sensitive to noise. By employing hierarchical knowledge propagation that leverages the complementary strengths of both types of features, our model significantly

TABLE 6
ABLATION STUDY ON THE EFFECT OF SELF-ATTENTION. (A) IDBHI DATASET. (B) PEDIATRIC DATASET

(a)

Model	Acc	Precision	Recall	F1-Score
w/ self-attention	**0.6597**	**0.6398**	**0.6118**	**0.6569**
w/o self-attention	0.5918	0.5668	0.4560	0.5054

(b)

Model	Acc	Precision	Recall	F1-Score
w/ self-attention	**0.8755**	0.8632	**0.8753**	**0.8690**
w/o self-attention	0.8485	**0.8756**	0.7991	0.8356

enhances classification performance to differentiate between normal and abnormal respiratory sounds [19]. To assess statistically significant improvements achieved by the proposed model, the Mann-Whitney U test was utilized. For the ICBHI dataset, the F1-score of the proposed model was significantly higher than that of models with 2 steps, 1 step, and no knowledge propagation, yielding p-values of 0.004 (< 0.05) for each comparison. Similarly, analysis of the Pediatric dataset produced p-values of 0.028 for both 2 steps and 1 step, and 0.004 (< 0.05) for the model without knowledge propagation. As shown in Table 5, the three-step propagation process achieved the highest accuracy and F1-score, with a better balance between precision and recall.

2) Effect of Self-Attention: In our model, we incorporated a self-attention operation for each extracted local feature prior to the knowledge propagation. As shown in Table 6, the omission of the self-attention module is associated with a significant degradation in performance. The statistical significance of our findings is affirmed by the Mann-Whitney U test, which produced a p-value of 0.004 (< 0.05) for both the ICBHI and Pediatric datasets. The absence of self-attention results in a substantial 15.58% and 7.62% reduction in recall for ICBHI and Pediatric dataset respectively. This result indicates the important role of self-attention as it selectively emphasizes the important information within each local feature by focusing on the most relevant features, which is required to be distilled and fused together for accurate classification. A lower value of recall is not favorable because the recognition of abnormal samples is crucial in the classification of respiratory sounds. Therefore, the integration of the self-attention mechanism constitutes an essential component of our proposed model, which enhances the effectiveness of feature distillation.

V. CONCLUSION

In this study, we have proposed an advanced DNN architecture specifically designed to automate respiratory sound classification. This innovation is particularly significant for the early diagnosis of respiratory diseases. The key element of our architecture is the knowledge propagation mechanism, which combines different levels of features to mitigate the potential information loss during the deep learning process. This knowledge propagation involves the strategic fusion of high-level information from the deepest convolutional layer with features extracted from previous layers. Through this process, the input knowledge is effectively propagated from high-level to low-level features. In addition, our model incorporates a self-attention mechanism to re-weight crucial features and channel average pooling to reduce feature dimensions while preserving essential time-frequency information. The experimental results confirm the effectiveness of our proposed model. The proposed method achieves a new state-of-the-art performance on the ICBHI benchmark dataset and also outperforms the RespireNet models on the self-collected Pediatric dataset. A notable strength of our model is its pronounced recall or sensitivity to accurately identify abnormal respiratory sounds, a critical requirement for effective disease diagnosis and management. These results confirm the potential of our approach as a robust tool in the field of respiratory health diagnostics.

Author Contributions: SWN, SKJ, and SK, conception of the project. SWN, SKJ, and SK, experimental protocol. IAPAC and SK, design of methodology. SWN and SKJ, data collection. IAPAC and SK, data analysis and software development. IAPAC and SK, drafting the manuscript. IAPAC, SWN, SKJ, and SK, review and editing. All authors contributed to the critical revision and finalisation of the manuscript.

Conflict of Interests: The authors declare no potential conflicts of interest.

REFERENCES

[1] GBD 2019 Chronic Respiratory Diseases Collaborators, "Global burden of chronic respiratory diseases and risk factors, 1990–2019: An update from the global burden of disease study 2019," *eClinicalMedicine*, vol. 59, 2023, Art. no. 101936.

[2] J. A. Dar, K.K. Srivastava, and A. Mishra, "Lung anomaly detection from respiratory sound database (sound signals)," *Comput. Biol. Med.*, vol. 164, 2023, Art. no. 107311.

[3] R. Zulfiqar, F. Majeed, R. Irfan, H. T. Rauf, E. Benkhelifa, and A. N. Belkacem, "Abnormal respiratory sounds classification using deep CNN through artificial noise addition," *Front. Med.*, vol. 8, 2021, Art. no. 714811.

[4] B. H, "The inventor of the stethoscope: René Laennec," *J Fam Pract.*, vol. 37, no. 2, 1993, Art. no. 191.

[5] A. Roguin, "Rene Theophile Hyacinthe Laennec (1781-1826): The man behind the stethoscope," *Clin. Med. Res.*, vol. 4, pp. 230–235, 2006.

[6] S. Leng, R. S. Tan, K. Chai, C. Wang, D. Ghista, and L. Zhong, "The electronic stethoscope," *Biomed. Eng. Online*, vol. 14, 2015, Art. no. 66.

[7] G. Serbes, S. Ulukaya, and Y. P. Kahya, "An automated lung sound preprocessing and classification system based on spectral analysis methods," *Precis. Medi. Powered pHealth Connected Health*, vol. 66, pp. 45–49, 2018.

[8] 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*, 2018, pp. 1–6.

[9] G. Serbes, C. O. Sakar, Y. Kahya, and N. Aydin, "Feature extraction using time-frequency/scale analysis and ensemble of feature sets for crackle detection," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2011, pp. 3314–3317.

[10] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. A. Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *Proc. Int. Conf. Artif. Neural Netw.*, 2018, pp. 208–217.

[11] Y. Kim et al., "Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning," *Sci. Rep.*, vol. 11, 2021, Art. no. 17186.

[12] T. L.-T. Niksa Jakovljevic, "Hidden Markov model based respiratory sound classification," *Precis. Med. Powered pHealth Connected Health*, vol. 66, pp. 39–43, 2018.

[13] S.-Y. Jung, C.-H. Liao, Y.-S. Wu, S.-M. Yuan, and C.-T. Sun, "Efficiently classifying lung sounds through depthwise separable CNN models with fused STFT and MFCC features," *Diagnostics*, vol. 11, 2021, Art. no. 732.

[14] 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 Inform.*, vol. 25, no. 7, pp. 2595–2603, Jul. 2021.

[15] L. Pham, H. Phan, R. Palaniappan, A. Mertins, and I. McLoughlin, "CNN-MoE based framework for classification of respiratory anomalies and lung disease detection," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 2938–2947, Aug. 2021.
[16] B. Ari, O. Alcin, and A. Sengur, "A lung sound classification system based on data augmenting using ELM-Wavelet-AE," *Biomed. Eng. Online*, vol. 17, pp. 79–88, 2022.
[17] S. Gairola, F. Tom, N. Kwatra, and M. Jain, "RespireNet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," in *Proc. IEEE 43rd Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2021, pp. 527–530.
[18] L. Xu, J. Cheng, J. Liu, H. Kuang, F. Wu, and J. Wang, "ARSC-Net: Adventitious respiratory sound classification network using parallel paths with channel-spatial attention," in *Proc. IEEE Int. Conf. Bioinform. Biomed.*, 2021, pp. 1125–1130.
[19] Y. Hua, L. Mou, and X. X. Zhu, "LAHNet: A convolutional neural network fusing low- and high-level features for aerial scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 4728–4731.
[20] G. Zhao, J. Wang, and Z. Zhang, "Random shifting for CNN: A solution to reduce information loss in down-sampling layers," in *Proc. Int. Jt. Conf. Artif. Intell.*, 2017, pp. 3476–3482.
[21] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Proc. Conf. Neural Inf. Process Syst.*, 2019, pp. 68–80.
[22] A. Sovijärvi, F. Dalmasso, J. Vanderschoot, L. Malmberg, G. Righini, and S. Stoneman, "Definition of terms for applications of respiratory sounds," *Eur. Respir. Rev.*, vol. 10, pp. 597–610, 2000.
[23] M. Sarkar, I. Madabhavi, N. Niranjan, and M. Dogra, "Auscultation of the respiratory system," *Ann. Thoracic Med.*, vol. 10, pp. 158–168, 2015.
[24] A. Bohadana, G. Izbicki, and S. Kraman, "Fundamentals of lung auscultation," *New England J. Med.*, vol. 370, pp. 744–751, 2014.
[25] H. Pasterkamp, S. Kraman, and G. Wodicka, "Respiratory sounds: Advances beyond the stethoscope," *Amer. J. Respir. Crit. Care Med.*, vol. 156, pp. 974–87, 1997.
[26] E. Andrés, R. Gass, A. Charloux, C. Brandt, and A. Hentzler, "Respiratory sound analysis in the era of evidence-based medicine and the world of medicine 2.0," *J. Med. Life*, vol. 11, pp. 89–106, 2018.
[27] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604–624, Feb. 2021.
[28] V. Roger, J. Farinas, and J. Pinquier, "Deep neural networks for automatic speech processing: A survey from large corpora to limited data," *Eurasip J. Audio Speech Music Process.*, vol. 2022, pp. 1–15, 2022.
[29] N. O'Mahony et al., "Deep learning vs. traditional computer vision," in *Proc. Conf. Adv. Comput. Vis.*, Springer, 2020, pp. 128–144.
[30] I. Moummad and N. Farrugia, "Pretraining respiratory sound representations using metadata and contrastive learning," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2022, pp. 1–5.
[31] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.
[32] E. Messner et al., "Multi-channel lung sound classification with convolutional recurrent neural networks," *Comput. Biol. Med.*, vol. 122, 2020, Art. no. 103831.
[33] 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.
[34] Z. Wang and Z. Wang, "A domain transfer based data augmentation method for automated respiratory classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 9017–9021.
[35] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 2736–2746.
[36] T. Nguyen and F. Pernkopf, "Lung sound classification using co-tuning and stochastic normalization," *IEEE. Trans. Biomed. Eng.*, vol. 69, no. 9, pp. 2872–2882, Sep. 2022.
[37] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.
[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.