Improved Lung Sound Classification Model Using Combined Residual Attention Network and Vision Transformer for Limited Dataset

Muhammad Jurej, Roslidar Roslidar, and Yunida Yunida
Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Indonesia

Article Info

Article history:

Received Apr 4, 2024
Revised Nov 26, 2024
Accepted Dec 28, 2024

Keywords:

Lung sound
ICBHI score
Residual attention network
Vision transformer

ABSTRACT

According to WHO data, the prevalence of respiratory disorders is increasing, exacerbated by a shortage of skilled medical professionals. Consequently, there is an urgent need for an automated lung sound classification system. Current methods rely on deep learning, but limited lung sound data resulted in low model accuracy. The widely used ICBHI 2017 dataset has an imbalanced class distribution, with a normal class at 52.8%, wheezing at 27.0%, crackles at 12.8%, and combined wheeze and crackles at 7.3%. The imbalance of the dataset may affect the model's efficiency and performance in classifying lung sounds. Given these data limitations, we propose a hybrid model, combining residual attention network (RAN) and vision transformer (ViT), to construct an effective respiratory sound classification model with a small dataset. We employ feature fusion techniques between convolutional neural network (CNN) feature maps and image patches to enrich lung sound features. Additionally, our preprocessing involves bandpass filtering, resampling sounds to 16 kHz, and normalizing volume to 15 dB. Our model achieves impressive ICBHI scores with 97.28% specificity, 92.83% sensitivity, and an average score of 95.05%, marking a 10% improvement over state-of-the-art models in previous research.

Corresponding Author:

Roslidar Roslidar,
Department of Electrical and Computer Engineering,
Universitas Syiah Kuala,
Jl. Tgk. Syech Abdurrauf, No.7 Darussalam, Kota Banda Aceh, Aceh 23127, Indonesia.
Email: roslidar@usk.ac.id

1. INTRODUCTION

In biology, respiration is a fundamental attribute shared by all living organisms, a facet prominently exemplified in mammals through the intricate mechanism of lung function. The lungs stand out as pivotal organs within the human physiological framework, as the World Health Organization (WHO) statistics show a staggering global health burden. With approximately 10 million individuals afflicted by tuberculosis (TB), 65 million grappling with chronic obstructive pulmonary disease (COPD), 334 million contending with asthma, and an alarming toll of 3 million lives succumbing to the grim specters of TB, lung cancer, and COPD, these maladies manifest as significant contributors to global mortality rates [1]. In light of these challenges, it is crucial to watch and assess the health of our lungs closely, identifying any issues that might affect breathing, including carefully listening to breath sounds using a stethoscope, which is a key method in lung check-ups [2].

Adventitious lung sounds (ASL), encompassing phenomena such as wheezes and crackles, stand as critical indicators of a spectrum of pulmonary disorders, including but not limited to asthma [3], COPD [4], interstitial lung disease [5], bronchiectasis [6], heart failure [7], and pneumonia [8]. The distinctive crackles manifest as explosive breathing sounds emanating from fluid bubbles within the trachea or bronchial tubes. In contrast, wheezing manifests as a high-pitched, whistle-like resonance resulting from the passage of air through constricted airways [9].

Currently, the challenge is that only healthcare professionals with training can pick up on these sounds. Meanwhile, the need for more specialists in this field adds a layer of complexity, potentially slowing down the ability to promptly identify the specific illness of one person with respiratory disorders. Automated examination of respiratory sounds can mitigate these limitations and facilitate the implementation of telemedicine applications for monitoring patients beyond traditional clinical settings, with the potential involvement of less-specialized personnel, such as community health workers.

The active investigation of algorithmic approaches for detecting lung disorders based on respiratory sounds has increased, especially with the advent of digital stethoscopes [10], [11]. Many studies on this subject focus on detecting aberrant respiratory sounds, such as wheezing and crackles. At the same time, initial studies emphasized the use of manually crafted features and conventional machine learning techniques [12]-[14]. In recent times, there has been a predominant focus on methods rooted in deep learning [4], [15], [16]; common features used to train deep learning models in lung sound detection systems are the Mel-spectrograms [17], [18], and The Mel-frequency cepstral coefficients (MFCCs) [19], [20] which are compiled into 2D data which is then trained with image level classification.

The deep learning models employed in prior research for lung sound classification included supervised learning convolutional neural networks (CNNs) [16], [21], [22], recurrent neural networks (RNNs) [20], [23], and hybrid models combining CNNs and Long Short-Term Memory (LSTM) [24]. Recent work conducted by Wu et al. [25] utilized feature fusion techniques that combine 3 features, including the spectrogram, the Mel spectrogram, and MFCCs, then utilized the CNN model with skip connection and resulted an ICBHI score of 88.5%. Research conducted by Mang et al. [26] used voice cochleogram features to improve time-frequency representation to optimize CNN; they got an ICBHI score of 85.1%. Bhushan et al. [27] proposed a CNN-LSTM self-attention model to classify respiratory sounds and got an ICBHI score of 57.02%. Then, Moummad et al. [28] used contrastive learning techniques to classify respiratory sounds. However, the implementation of the vision transformer (ViT) model in previous research still needs to be improved. The use of transformer models for audio classification tasks, namely the audio spectrogram transformer (AST) [29], has been carried out with satisfactory results. Ariyanti et al. [30] used the AST model for respiratory sound classification and utilized the Mel spectrogram feature, and they got an ICBHI score of 69.3%.

ViT has a larger receptive field compared to CNNs, which generally use 3×3 kernels, which are limited to only local neighborhood capture local representation because the size of the receptive field is very important to construct a contextual visual understanding [31]. The integration of CNNs and ViT architectures is driven by the intention to harness the complementary strengths of both models. CNNs are recognized for their efficiency in capturing local patterns and spatial hierarchies, particularly in smaller images, due to their translation equivariance and weight-sharing properties. On the other hand, transformers, exemplified by ViT, come with multi-head attention, and excel in modeling long-range dependencies but are often deemed data-hungry, presenting challenges in scenarios with limited datasets. The combination aims to create a hybrid model that can generalize well on smaller datasets, providing a flexible and efficient solution that balances the strengths of CNNs and transformers [32]. In a prior study conducted by J. Neto et al. [33], the combination of CNN and ViT was explored using a dataset comprising breathing sounds. In their research, they integrated a convolutional block attention module (CBAM) as the convolutional block. They adopted data-efficient image transformers (DeIT) as the transformer block, and the research still gets an ICBHI score that is still unsatisfactory at 57.36%.

In this paper, we propose a lung sound classification model using a combining network of RAN and ViT. RAN has the ability to generate attention features by utilizing the soft mask branch and trunk branch mechanisms. The mask branch consists of two main processes, namely fast feed-forward sweeps and top-down feedback. The feed-forward step quickly collects overall information from the entire image, while the top-down feedback step integrates the global information with the initial feature map. With this mechanism, RAN is able to generate low-noise features. RAN also has better capabilities in the case of image net data classification, where RAN with attention-92 settings gets a top-error of 19.5% and a top-5 error of 4.8% [34], compared with CBAM with the ReNetXt101+CBAM combination setting, the top-1 error was 21.07%, and the top-5 error was 5.59 [35].

Works on developing lung sound classification model implementing the deep learning models used the ICBHI 2017 dataset [36], [37], which consists of 6,898 respiratory cycles consisting of 1,864 wheezing sounds, 886 crackles, 506 combinations of wheezing and crackles, and 3,642 normal sounds. There needs to be more than this dataset to train a deep learning network. Therefore, in this work, we also enriched the dataset by using simple audio augmentations such as pitch shifting, time shifting, time stretching, and pitch stretching. Then, this dataset was fed to the proposed network, using a combination of RAN and ViT to allow more effective utilization of the limited lung sound dataset.

The main contributions of our work are:

- Proposing a fusion technique of lung sound features between the CNN feature map and image patch to enrich features.
- Proposing a hybrid model that results from training in combined RAN and ViT to build an effective respiratory sound classification model with a limited dataset.
- Demonstrating RAN-ViT performance on the ICBHI 2017 dataset using specificity and score metrics that outperform other state-of-the-art approaches.

The rest of this article describes the dataset preparation and proposed network in Section 2. Then, the simulation result and model performance are discussed in Section 3. Finally, Section 4 sums up the findings.

2. METHOD

This section describes the respiratory sound dataset obtained from the ICBHI 2017, the proposed model by combining RAN and ViT, the training setting, and model validation. The discussion is clearly described in the following subsections.

2.1. Dataset Preparation

In this study, we used the ICBHI 2017 dataset [36], [37], which is a very popular dataset for building machine learning and deep learning models capable of classifying respiratory sounds. The dataset contains 6,898 breathing cycles, including 3,642 normal sounds, 1,864 crackles, 886 wheezing sounds, and 506 sounds of both (crackles and wheezing) consisting of a total of 5.5 hours of recordings; the number of audio samples consists of 920 recordings and 126 patients. Figure 1 shows the distribution of imbalance for each class in this ICBHI dataset.

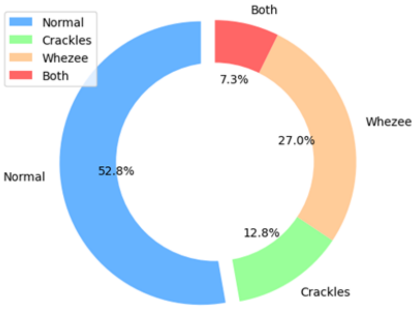


Figure 1. ICBHI dataset class distributions

There are two types of files in the ICBHI dataset, namely files with .wav and .txt extensions. The file with the .wav extension contains a recording of the patient's breathing sounds, while the .txt file contains an annotation of the breathing sounds. Each respiratory sound file has an annotation file that indicates whether the sound is in the category of wheeze, crack, both, or normal. This annotation file has four columns, namely the start time column (the time the breath starts), the end time column (the time the breath ends), the crack column (a crack sound marker with a value of 1 and 0 if there is no crack sound), and the is wheeze column (wheeze sound marker with a value of 1, and 0 if there is no wheeze sound). For example, Figure 2 shows the data preparation process before undergoing the data preprocessing process. In this process, the sound in the .wav file will be splitted and divided according to the annotation in the .txt file.

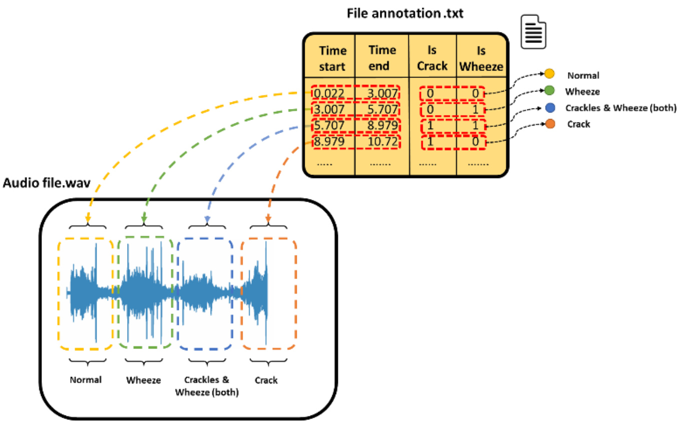


Figure 2. ICBHI dataset preparation process

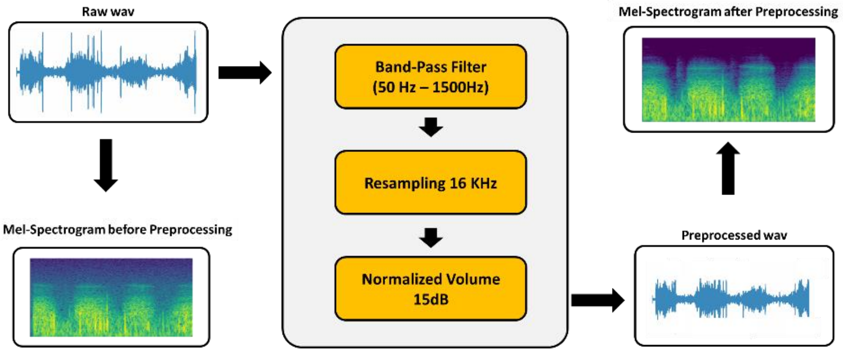

Raw wav
Band-Pass Filter (50 Hz – 1500Hz)
Resampling 16 KHz
Normalized Volume 15dB
Mel-Spectrogram after Preprocessing
Mel-Spectrogram before Preprocessing
Preprocessed wav

Figure 3. ICBHI dataset preprocessing process

As illustrated in Figure 2, for the example of a breathing start time of 0s and a breathing end time of 3s. If the crack and wheeze sound values are 0 or there are no crack and wheeze sounds during breathing, then the sound is categorized as a normal sound class. If the crack sound is 0 and the wheeze sound is 1, then the sound is classified as the wheezing sound class. Conversely, if the crack sound is 1 and the wheeze sound is 0, it is classified as the crack sound class. If both crack and wheeze have a value of 1, then they are categorized as both sound classes. After the sounds are splitted based on their annotations, the files will be saved in the folder corresponding to each sound class.

After the data preparation process, the next step is the preprocessing process. This ICBHI dataset has a non-uniform sampling rate and power volume and still has a lot of noise, such as people talking, alarm sounds, and other objects. Therefore, we do some preprocessing that makes the data more uniform and low noise. Preprocessing to achieve consistency, we harmonize the sampling rates of the dataset recordings, which initially range from 4 kHz to 44.1 kHz, by setting all recordings to 16 kHz and setting the volume of all recordings to 15 dB. To handle the noise in the recording, we used a bandpass filter with a low cutoff of 50 Hz and a high cutoff of 1500 Hz. The detailed process at this preprocessing stage is shown in Figure 3.

2.2. Proposed Model

We propose a model that can produce training models with good results even though the dataset is small and unbalanced. In this study, we created a CNN-Transformer hybrid model, which can take advantage of the advantages of each model, as shown in Figure 4. The CNN model of RAN has the ability to produce low-noise feature maps and attention features, and it can capture local neighborhood context [34]. The Residual Attention Network is composed of stacked Attention Modules, each featuring two distinct branches: the mask branch and the trunk branch. The mask branch generates attention masks to emphasize critical features and suppress less relevant ones, while the trunk branch processes feature and seamlessly integrates with any modern network architecture, enhancing its flexibility and performance [34]. Pre-activation Residual Units, ResNeXt, and Inception are utilized as foundational of RAN, the trunk branch processes the input x to produce an output $\boldsymbol{T}(\boldsymbol{x})$, while the mask branch employs a bottom-up top-down structure to generate a mask $\boldsymbol{M}(\boldsymbol{x})$ of the same size. This mask softly weights the output features $\boldsymbol{T}(\boldsymbol{x})$, mimicking the feedforward and feedback attention process. Neuron control gates are implemented similarly to Highway Networks, resulting in the final module output. The output of Attention Module of RAN $(\boldsymbol{H})$ is:

$$\boldsymbol{H}_{\boldsymbol{i},\boldsymbol{c}}(\boldsymbol{x}) = M_{i,c}(x) * T_{i,c}(x) \tag{1}$$

where *i* ranges over all spatial positions and $c \in \{1, \dots, C\}$ is the index of the channel.

In Attention Modules, the attention mask functions not only as a feature selector during forward inference but also as a gradient filter during backpropagation. Within the soft mask branch, the gradient of the mask with respect to the input features is expressed as.

$$\frac{\partial M\,(x,\theta)T(x,\ \emptyset)}{\partial \emptyset} = M(x,\theta)\frac{\partial T\,(x,\ \emptyset)}{\partial \emptyset} \tag{2}$$

Where θ represents the parameters of the mask branch, and $\emptyset$ corresponds to the parameters of the trunk branch. This characteristic enhances the robustness of Attention Modules against noisy labels, as the mask branch helps prevent incorrect gradients, caused by noisy labels, from affecting the updates of the trunk branch parameters.

Transformer model that we use on this work captures global neighborhood context using ViT, which is successful with the AST [29] model, which produces good accuracy for audio data classification, which has the ability to multi-head attention and a large receptive field so that it can capture global context on data features. The self-attention mechanism is a crucial component of Transformers, designed to explicitly capture interactions among all elements within a sequence. In this context, the sequence refers to the patches of an image mel spectrogram, which has been divided into 16 × 16 segments. Essentially, a self-attention layer updates each element of a sequence by gathering global information from the entire input sequence. Let's denote a sequence of *n* entities $(x_1, x_2, \ldots, x_n)$ by $\boldsymbol{X} \in R^{n\times d}$ where *d* is the embedding dimension to represent each patch image (sequence entities). The purpose of self-attention is to capture the interactions between all *n* entities by representing each entity based on the global contextual information. This is achieved by introducing three learnable weight matrices to transform the Queries $(\boldsymbol{W}^Q \in R^{n\times d_q})$, Keys $(\boldsymbol{W}^K \in R^{n\times d_k})$, and Values $(\boldsymbol{W}^V \in R^{n\times d_v})$, where $d_q = d_k$ **.** The input sequence $\boldsymbol{X}$ is f irst projected onto these weight matrices to get $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^k$ and $\boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}^v$. The output $\boldsymbol{Z} \in R^{n\times d_v}$ of the self attention layer is:

$$\boldsymbol{Z} = softmax\left(\frac{QK^T}{\sqrt{d_q}}\right) \tag{3}$$

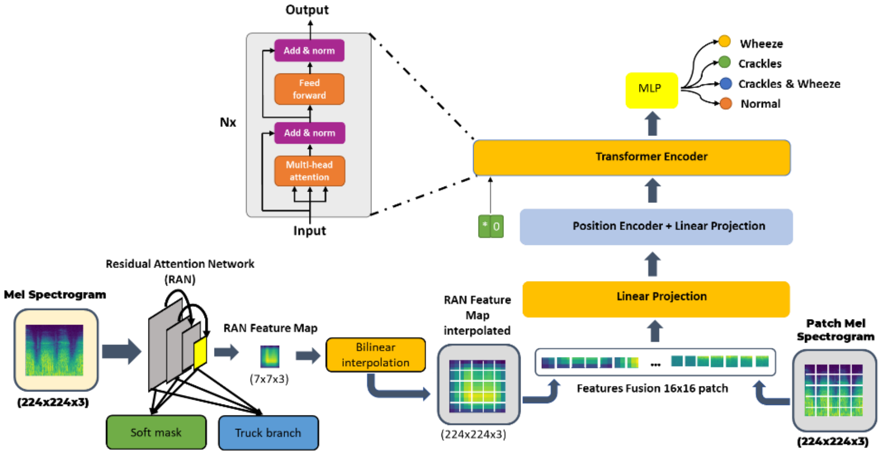


Figure 4. Feature fusion technique to combine RAN and ViT

The first process carried out in the model we propose is converting data that has been previously processed by removing some noise and resampling it into data in the form of a Mel spectrogram representation. Mel spectrogram is a technique based on the human sensory system and applied to the depiction of time-frequency audio input [38]. We use a window size setting for the FFT of 2048 and an overlap of 128. We use the help of the Librosa framework to generate a Mel spectrogram image from the representation of the respiratory sound signal that has been previously preprocessed. We resized the Mel spectrogram to 224×224 pixels because the input for RAN is 224×224, and the RAN model setting was RAN attention-92.

The previously produced Mel-spectrogram image then undergoes a feature fusion process consisting of two stages. In the first stage, the Mel-spectrogram image is fed into the RAN attention92 model, which produces a feature map with a size of 7×7. The proposed RAN model had the fully connected layer removed. The resulting feature map is then enlarged using bilinear interpolation techniques to the original Mel-spectrogram size. The interpolation results are then broken down into patches with a size of 16×16 pixels, producing 196 image tokens. In the second stage, the Mel-spectrogram image is directly cut into the same patch size as in the previous process without going through the RAN model. It also generates 196 mel-spectrogram image tokens. The two sets of tokens, each from the two stages, are then combined by concatenation to form 392 tokens. This combination then becomes input for the ViT model in the classification process. The details of this process regarding feature fusion are explained in Figure 5.

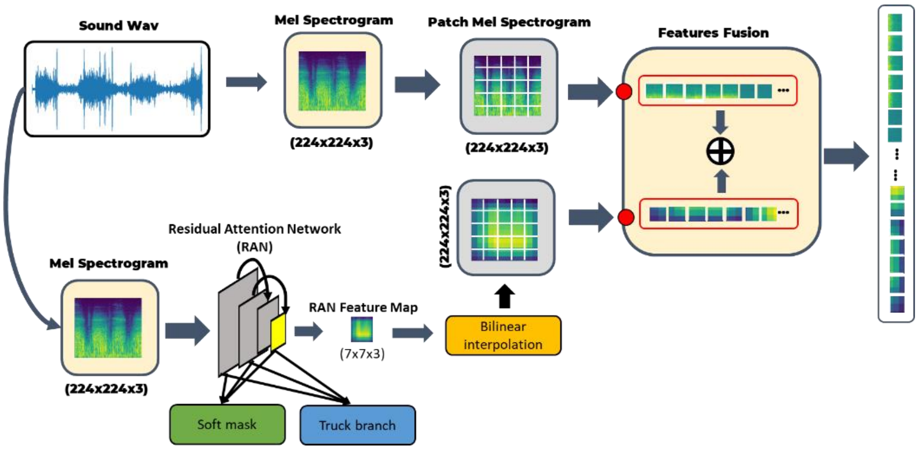


Figure 5. Overview of the proposed model RAN and ViT combination

The previously diffused features resulted in 392 tokens, which will serve as input to be fed into the ViT model. The ViT model we use in this study is the vanilla ViT version with customized configurations, including a head attention count of 8, multi-layer perceptron (MLP) dimension of 1024, linear projection dimension of 1024, and head dimension of 64. The 392 tokens are then piped into the encoder transformer of the ViT model, which then undergoes a classification stage in the last MLP layer. This process allows the ViT model to understand and extract complex features from the token representations given earlier, leading to more accurate decision-making in the classification stage.

2.3. Training Setting

The training process was conducted in sequential steps to ensure reproducibility and efficiency, utilizing a dual-GPU configuration. First, the PyTorch Lightning framework was used to implement the model, leveraging the DistributedDataParallel (DDP) strategy for synchronous training across two GPUs. This approach efficiently distributed computations and gradients, enabling faster training times and scalability. The specific steps involved in the training process are as follows:

1. Hardware Setup: Two RTX 3070 GPUs with 8 GB of VRAM each were used to accelerate the training process. RAM 32 GB, CPU Intel i7.
2. Batch Size Configuration: A batch size of 32 was selected to maximize GPU memory utilization without compromising computational efficiency.
3. Optimizer Selection: The AdamW optimizer was employed, with three variations of the learning rate (0.0001, 0.003, and 0.005) to identify the optimal value for stable learning.
4. Learning Rate Scheduling: The StepLR scheduler was applied, with a step size of 20 epochs and a gamma factor of 0.2. This adjustment gradually reduced the learning rate, ensuring improved convergence and training stability.
5. Training Duration: The model was trained for a total of 150 epochs, enabling the optimizer to reach convergence without overfitting.
6. Loss Function: Cross-entropy loss was used to handle the multi-class classification task, ensuring proper optimization of the model's predictions.

For more detail about implementation on training proses, the following is the psudocode the training process as show on Algorithm 1.

Algorithm 1: Training Process

```
Input: Dataset D, batch size N, LightningModule M, optimizer O, trainer T, accelerator
(GPU), devices (2), RAN x ViT architecture, Mel spectrogram representation, epochs=150

Output: Trained model M.
Initialize Environment
Preprocess Data:
    1. Remove Noise and Resample data.
    2. Convert data to Mel spectrogram representation:
        Window size for FFT = 2048, overlap = 128
        Resize Mel spectrogram to 224×224 pixels using Librosa.
Define RAN x ViT Architecture:
    1. Stage 1: Process through RAN Attention-92:
        Input Mel spectrogram to RAN → Generate 7×7 feature map
        Remove fully connected layer in RAN
        Upscale feature map to 224×224 using bilinear interpolation
        Split resized feature map into 16×16 patches → Generate 196 tokens
    2. Stage 2: Direct Tokenization of Mel spectrogram:
        Split original Mel spectrogram into 16×16 patches → Generate 196 tokens
    3. Feature Fusion:
        Concatenate tokens from Stage 1 and Stage 2 → Total 392 tokens
    4. ViT Encoder:
        Input 392 tokens to ViT encoder with the following configurations:
            Head attention count = 8, MLP dimension = 1024, linear projection =
            1024, head dimension = 64
    5. Classification: Use last MLP layer in ViT for final classification.
Detail Train Model Process (vanila):
    for k in {1, 2, …, epochs} do:
        for i in {1, …, batches in D} do:
            x, y ← Sample batch from D
            y_pred ← M.forward(x)
            loss ← Compute Cross-Entropy Loss(y_pred, y)
            Backpropagation and optimization step
        end for
    end for
Initialize Trainer (Pytorch lighning module): T ← Trainer(accelerator='gpu',
devices=2, max_epochs=epochs, strategu:"ddp")
Train Model: T.fit(M, D)
return trained model M.
```

2.4. Model Validation

We evaluated our model using the ICBHI dataset [32], in which our model classifies 4 classes of breathing sounds provided in the dataset, namely normal, crackles, wheeze, and both. We divided the dataset with 60% training data settings, and 40% testing data, then we evaluated the results of the model performance using the metrics commonly used on the ICBHI dataset from eq (1), eq. (2), and eq. (3).

$$S_e = \frac{P_c + P_w + P_b}{N_c + N_w + N_b} \tag{4}$$

where S_e is the ICBHI sensitivity score, P_c, P_w, and P_b respectively is the number of crackles, wheeze, and both classes correctly classified, and N_c, N_w and N_b is total number of samples of crackles, wheeze, and both classes, respectively.

$$S_p = \frac{P_n}{N_n} \tag{5}$$

where S_p is the ICBHI specificity score, P_n is number of normal classes correctly classified, and N_n is total number of normal classes.

$$S_c = \frac{S_e + S_p}{2} \tag{6}$$

where S_c is the average ICBHI score.

3. RESULTS AND DISCUSSION

In this section, we discuss the results of each process step to achieve the results and performance of our proposed model. The first step is preprocessing data that has been prepared previously and has been split according to the annotation. The second is training the network. And lastly testing the model performance.

3.1. Preprocessing Result

This process involves using a preprocessing pipeline with a band-pass filter to remove noise from the sound, then resamples the sound to 16 kHz, and finally normalizes the volume to 15 dB. The results of the sound samples that have gone through this preprocessing stage produce a Mel-spectrogram that is clean from noise, as shown in Figure 6.

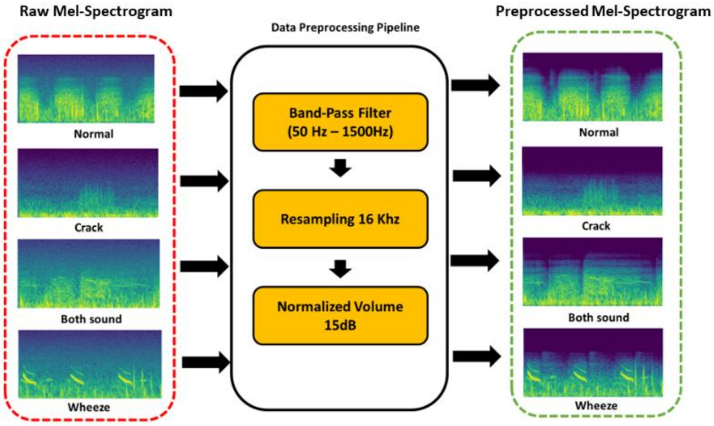


Figure 6. Dataset preprocessing pipeline

In the audio processing, a Mel-spectrogram image is initially generated, ensuring a clean representation by eliminating noise. This image is then resized to a standardized size of 224×224 pixels before being fed into a RAN model. The RAN model plays a crucial role in extracting high-level features by focusing on patterns indicative of specific sounds. The resulting feature map produced by the RAN serves as a refined representation, highlighting dominant frequencies and key characteristics that facilitate easier classification of distinct sounds. The generated Mel-spectrogram image encompasses a frequency range from 0 to 8,192 Hz, allowing the network to distinguish the distinctive characteristics of the dominant frequency associated with each sound. The intensifying yellow hues in the image signify the presence of a highly prominent frequency at specific points in time. According to the American Thoracic Society [39], distinct frequencies characterize different ASL. Wheezing sounds, for instance, exhibit a frequency range of 400–2,500 Hz throughout 80 µs. In contrast, crackle sounds manifest at 60–350 Hz within a shorter time frame of 15 µs than wheezing sounds. On the other hand, normal lung sounds fall within the frequency range of 200–800 Hz, as outlined by the same authoritative source.

Figure 7 displays the results of the feature map generated with RAN. This feature map reflects the characteristics of each type of sound. In normal speech, the feature map gives more weight to the frequency range (128–512 Hz). Likewise, in crack sounds, the feature map is focused on low frequencies (60–256 Hz), according to the appearance of crack sounds at that frequency, as seen in the feature map image. Meanwhile, the wheezing sound is focused on the frequency range (256–1,024 Hz).

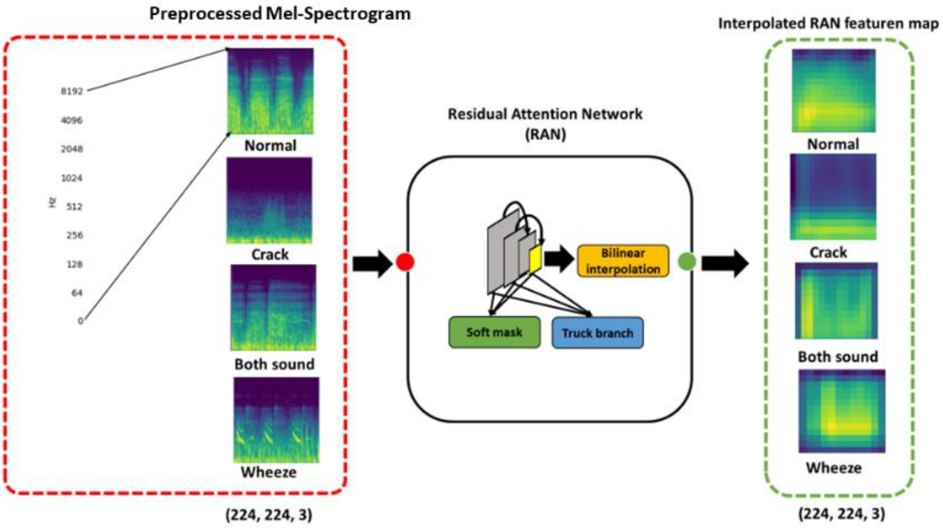
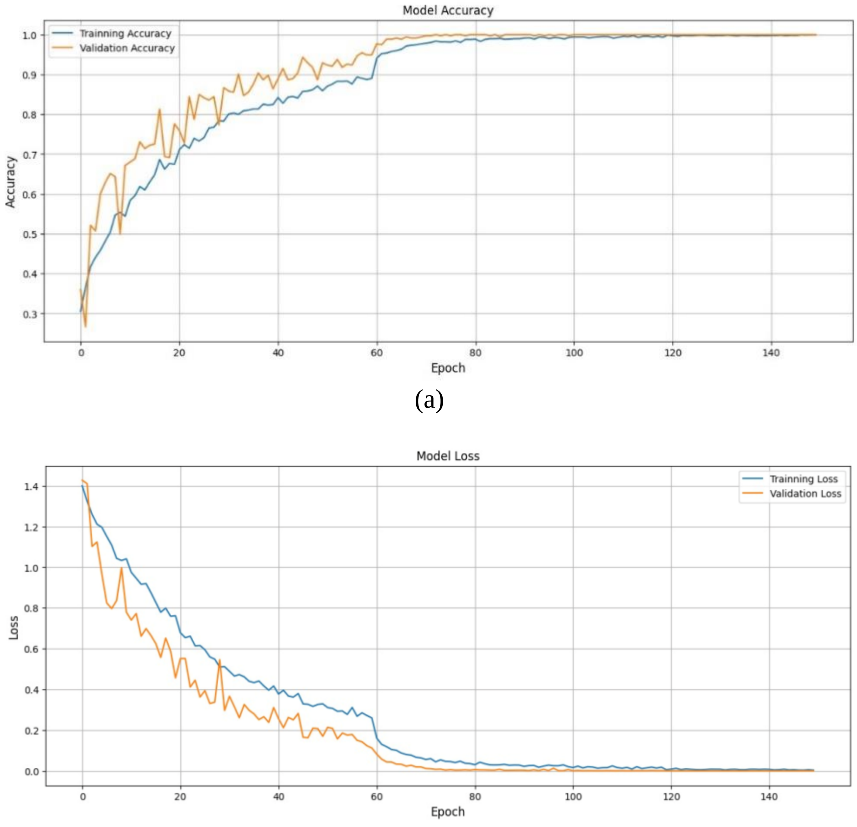


Figure 7. RAN features map after interpolation process

3.2. Network Performance

The features generated by RAN are divided into image patches with a size of 16×16 pixels. These patches are then merged with Mel-spectrogram images patched with the same size, namely 16×16 pixels. As a result, 28 image tokens are formed, which are subsequently input into the transformer encoder. This encoder is then trained with the settings explained in the training configuration above. Testing is conducted by trying three learning rate levels, with other settings remaining constant, such as the AdamW optimizer, cross-entropy loss, batch size of 32, and 150 epochs. The learning performance is shown in the learning curve in Figure 8.



(a)

(b)

Figure 8. Learning curve model: (a) training accuracy and (b) training loss

The learning curves in Figure 8 show the best performance of the combined network of RAN and ViT with hyperparameters of learning rate 0.0001, a training accuracy rate of 99.91 %, and a validation accuracy rate of 99.17%. The training curve shows that the model has a stable performance and no fluctuating values. Combining RAN with ViT and utilizing the feature fusion technique performs better when trained on a learning rate of 0.0001. The entire training process required a total time of 31,152.68 seconds (~8.65 hours) when utilizing the dual-GPU configuration. The trained model consists of 105,435,350 parameters and produces a model size of 988.3 MB. The average GPU utilization during training was 85 %, peaking at 93%. CPU utilization averaged 30%, peaking at 45%, and RAM utilization averaged 15.67 GB. For inference, which was performed on 32 samples, the total time required was 27.8 seconds. The GPU utilization during inference averaged 50%, peaking at 60%.

Our proposed model excels in capturing both local and global contexts within the Mel-spectrogram features of respiratory sounds. The RAN model production of low-noise map features ands attention filters, followed by their fusion with the Mel-spectrogram patch, enhances the guidance of the ViT model. It allows the enriched feature set, which includes the Mel-spectrogram patch, to integrate features from RAN. Given ViT reputation as a data-hungry model, the feature-level fusion system significantly improves the ViT model performance and reliability, especially when trained with a relatively limited dataset from ICBHI.

3.3. Model Testing

We validated our model using the ICBHI metric of specificity, sensitivity, and an average score, respectively, S_p, S_e, and S_c. We used the testing data and got the highest rate at S_p of 97.28%, S_e of 92.83%, and of S_c 95.05%. From the confusion metric in Figure 9, the proposed model can classify each class in the testing data well. We also compared previous research's state-of-the-art (SOTA) models, as provided in Table 1. Based on the comparison, the model we are proposing has an increase of 10% in terms of S_c, which is higher than other SOTA models.

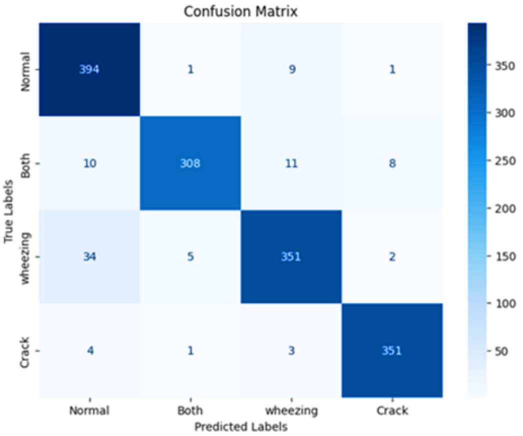


Figure 9. Model confusion matrix.

Table 1. The performance of the proposed model with other the state-of-the-art models

Method	ICBHI Scores		
	S_e (%)	S_p (%)	S_c (%)
DeIT + CBAM [33]	36.41	78.31	57.36
CNN+CBA+BRC+FT [16]	40.1	72.3	56.2
CNN-LSTM + Focal Loss [24]	60.29	84.26	68.52
AST [29]	52.1	86.4	69.3
M-SCL [28]	82.24	88.62	85.43
Cochleogram features + VGG16 [26]	53.45	68.71	61.08
Proposed model	**92.83**	**97.28**	**95.05**

To evaluate the performance of the proposed model more comprehensively, we assess it using class-wise metrics, including sensitivity, precision, and F1-score. These metrics provide a detailed analysis of the model's ability to correctly predict each class, as well as its robustness in handling imbalanced class distributions. The confusion matrix results were analyzed, and the class-wise metrics are presented in Table 2.

Table 2. Class-wise Metrics for the Proposed Model

Class	Metrics		
	$Sensitivity$ (%)	$Precision$ (%)	$F1-Score$ (%)
Normal	97.28	89.14	93.03
Both	91.39	97.78	94.48
Wheezing	89.54	93.85	91.64
Crack	97.77	96.96	97.36

The sensitivity (or recall) measures the model ability to correctly identify all positive instances of a particular class, with the 'Normal' class achieving 97.28%, indicating that most 'Normal' instances were correctly predicted. Precision evaluates the model's capability to avoid false positives by predicting a class only when it is confident, with the 'Both' class recording the highest precision at 97.78%, reflecting the model's strong ability to avoid misclassifications in this category. F1-score, the harmonic mean of sensitivity and precision, provides a balanced measure of the model's accuracy, where the 'Crack' class achieved the highest F1-score at 97.36%, demonstrating exceptional performance in identifying and classifying this class correctly. The overall results show that the proposed model performs consistently well across all classes, particularly excelling in detecting 'Normal' and 'Crack' classes with high sensitivity and precision. However, the slightly lower sensitivity for the 'Wheezing' class (89.54%) highlights room for improvement in detecting all instances of this class. Overall, the proposed model effectively handles the class imbalances present in the dataset and demonstrates strong generalizability for respiratory sound classification tasks.

4. CONCLUSION

This article introduces a hybrid model that combines the strengths algorithm of RAN and ViT. Notably, this hybrid model demonstrates effectiveness even when working with limited lung sound datasets. Our proposed model establishes a new state-of-the-art (SOTA) RAN and ViT combination network trained and tested on the ICBHI dataset for 4-class classification. The model shows outstanding performance with S_e of 92.83%, S_p of 97.28%, and S_c of 95.05%, showcasing a notable improvement of 10% over existing works. The proposed hybrid model combining Residual Attention Network (RAN) and Vision Transformer (ViT) demonstrates strong performance but has some limitations. The model is very complex and large, which increases computational demands, impacts training and inference time, and may limit deployment on resource-constrained devices. Additionally, the small and imbalanced dataset may affect its generalizability, and the sensitivity for the "Wheezing" class is slightly lower than for other classes. Future work could focus on creating a more efficient and lightweight model, using larger and more diverse datasets, optimizing the model for real-time use, and improving preprocessing techniques. Integrating additional data sources to enhance accuracy and applicability, along with efforts to make the model more interpretable, would also be beneficial.

REFERENCES

[1] Forum of International Respiratory Societies, *The Global Impact of Respiratory Disease*, 2nd ed. Sheffi eld, European Respiratory Society, 2017.

[2] A. M. Alqudah, S. Qazan, and Y. M. Obeidat, “Deep learning models for detecting respiratory pathologies from raw lung auscultation sounds,” *Soft comput.*, vol. 26, no. 24, pp. 13405–13429, 2022.

[3] T. Aptekarev, V. Sokolovsky, E. Furman, N. Kalinina, and G. Furman, “Application of deep learning for bronchial asthma diagnostics using respiratory sound recordings,” *PeerJ Comput Sci.*, vol. 9, pp. e1173, 2023.

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

[5] N. Klaembt, R. Conradt, U. Koehler, W. Cassel, and P. Fischer, “Overnight registration of crackles, cough and wheezing in patients with interstitial lung disease”, *A. miner*, 2023.

[6] B. Herrero-Cortina, M. Francín-Gallego, A. Sáez-Pérez, J., M. San Miguel-Pagola, L. Anoro-Abenoza, C. Gómez-González, J. Montero-Marco, M. Charlo-Bernardos, E. Altarribas-Bolsa, A. Pérez-Trullén, and C. Jácome, “Reliability and Validity of Computerized Adventitious Respiratory Sounds in People with Bronchiectasis,” *J Clin Med*, vol. 11, no. 24, 2022.

[7] H. Melbye, J. Ravn, M. Pabiszczak, L. A. Bongo, J. Carlos, and A. Solis, “Validity of deep learning algorithms for detecting wheezes and crackles from lung sound recordings in adults,” *medRxiv*, vol. 75, pp. 1–12, 2022.

[8] J. C. Aviles-Solis, C. Jácome, A. Davidsen, R. Einarsen, S. Vanbelle, H. Pasterkamp, and H. Melbye, “Prevalence and clinical associations of wheezes and crackles in the general population: The Tromsø study,” *BMC Pulm Med*, vol. 19, no. 1, pp. 1–11, 2019.

[9] J. S. Park, K. Kim, J. H. Kim, Y. J. Choi, K. Kim, and D. I. Suh, “A machine learning approach to the development and prospective evaluation of a pediatric lung sound classification model,” *Sci Rep*, vol. 13, no. 1, pp. 1–10, 2023.

[10] S. Reichert, R. Gass, C. Brandt, and E. Andrès, “Analysis of Respiratory Sounds: State of the Art,” *Clinic. Med.: Circul., Respir. and Pulmon. Med*., vol. 2, no. 5, 2008.

[11] H. Useyin Polat, Inan, and Guler, “A Simple Computer-Based Measurement and Analysis System of Pulmonary Auscultation Sounds,” *J. of Med. Syst*., vol. 28, pp. 665–672, 2005.

[12] N. S. Haider, B. K. Singh, R. Periyasamy, and A. K. Behera, “Respiratory Sound Based Classification of Chronic Obstructive Pulmonary Disease: A Risk Stratification Approach in Machine Learning Paradigm,” *J Med Syst*, vol. 43, no. 8, 2019.

[13] N. Jakovljević, *et al*., “Hidden Markov model based respiratory sound classification,” in *International Federation for Medical and Biological Engineering, 2009. IFMBE Proceedings*, 2009. *Springer Verlag*, 2018, pp. 39–43.

[14] G. Chambres, *et al*., “Automatic Detection of Patient with Respiratory Diseases Using Lung Sound Analysis,” in *2018 International Conference on Content-Based Multimedia Indexing (CBMI), IEEE*, 2018, pp. 1–6.

[15] 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, 2020.

[16] S. Gairola, *et al.,* “RespireNet: A Deep Neural Network for Accurately Detecting Abnormal Lung Sounds in Limited Data Setting,” in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2021, pp. 527-530.

[17] R. Liu, *et al*., “Detection of Adventitious Respiratory Sounds based on Convolutional Neural Network,” in *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), IEEE*, 2019, pp. 298–303.

[18] L. Shi, K. Du, C. Zhang, H. Ma, and W. Yan, “Lung Sound Recognition Algorithm Based on VGGish-BiGRU,” *IEEE Access*, vol. 7, pp. 139438–139449, 2019.

[19] J. J. M. Escobar, O. Morales Matamoros, R. Tejeida Padilla, L. Chanona Hernández, J. P. F. Posadas Durán, A. K. Pérez Martínez, I. Lina Reyes, and H. Quintana Espinosa, “Biomedical signal acquisition using sensors under the paradigm of parallel computing,” *Sensors (Switzerland)*, vol. 20, no. 23, pp. 1–36, 2020.

[20] K. Kochetov, *et al*., “Noise masking recurrent neural network for respiratory sound classification,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2018, pp. 208–217.

[21] 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, 2022.

[22] F. Demir, A. M. Ismael, and A. Sengur, “Classification of Lung Sounds with CNN Model Using Parallel Pooling Structure,” *IEEE Access*, vol. 8, pp. 105376–105383, 2020.

[23] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F. Maria, S. Jüttner, H. Olschewski, and F. Pernkopf, “Multi-channel lung sound classification with convolutional recurrent neural networks,” *Comput Biol Med*., vol. 122, pp. 103831, 2020.

[24] G. Petmezas, G. A. Cheimariotis, L. Stefanopoulos, L. Rocha, R. P. Paiva, A. K. Katsaggelos, and N. Maglaveras, “Automated Lung Sound Classification Using a Hybrid CNN-LSTM Network and Focal Loss Function,” *Sensors*, vol. 22, no. 3, 2022.

[25] C. Wu, D. Lei, *et al*., “Respiratory Disease Classification Model Based on Feature Fusion,” in *2023 4th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, IEEE, Aug. 2023, pp. 148–155.

[26] L. D. Mang, F. J. Canadas-Quesada, J. J. Carabias-Orti, E. F. Combarro, and J. Ranilla, “Cochleogram-based adventitious sounds classification using convolutional neural networks,” *Biomed Signal Process Control*, vol. 82, pp. 104555, 2023.

[27] P. Bhushan *et al*., “A Self-Attention Based Hybrid CNN-LSTM Architecture for Respiratory Sound Classification,” *GMSARN International Journal*, vol 18, pp. 54-61, 2024.

[28] I. Moummad, *et al*., “Pretraining Respiratory Sound Representations using Metadata and Contrastive Learning,” in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2023, pp. 1-5.

[29] Y. Gong, Y.-A. Chung, and J. Glass, “AST: Audio Spectrogram Transformer,” *arXiv*, vol. 3, 2021.

[30] W. Ariyanti, *et al*., “Abnormal Respiratory Sound Identification Using Audio-Spectrogram Vision Transformer,” in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2023, pp. 1–4.

[31] H. Yan, Z. Li, W. Li, C. Wang, M. Wu, and C. Zhang, “ConTNet: Why not use convolution and transformer at the same time?,” *arXiv*, vol. 3, 2021.

[32] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, “Escaping the Big Data Paradigm with Compact Transformers,” *arXiv*, vol. 4, 2022.

[33] J. Neto, *et al*., “Convolution-Vision Transformer for Automatic Lung Sound Classification,” in *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Natal, Brazil, 2022, pp. 97-102.

[34] F. Wang *et al*., “Residual Attention Network for Image Classification,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA: IEEE, Jul 2017, pp. 6450–6458.

[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” *Computer Vision – ECCV*, vol. 11211, pp. 3–19, 2018.

[36] B. M. Rocha *et al*., “A respiratory sound database for the development of automated classification,” in *IFMBE Proceedings*, Springer Verlag, 2018, pp. 33–37.

[37] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, N. Maglaveras, R. Pedro Paiva, I. Chouvarda, and P. de Carvalho, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiol Meas*, vol. 40, no. 3, 2019.

[38] B. Ustubioglu, G. Tahaoglu, and G. Ulutas, "Detection of audio copy-move-forgery with novel feature matching on Mel spectrogram," *Expert Syst Appl*, vol. 213, p. 118963, 2023.

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

BIOGRAPHY OF AUTHORS

Muhammad Jurej Alhamdi received his bachelor degree from the Department of Electrical Engineering, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh in 2022. He was a Research Assistant at the control system Laboratory in 2020-2021. His research interests include creating and designing autonomous car using Deep Learning, and application of deep learning in the world of health, a. Currently, he is pursuing her study in Master of Electrical Engineering, Universitas Syiah Kuala

Roslidar Roslidar received her Bachelor Degree in Electrical Engineering in 2001 from Universitas Syiah Kuala. In 2009 she graduated from the Master program in Telecommunication Engineering, University of Arkansas, USA, under Fulbright scholarship. In January 2022, she received her PhD in Doctoral School of Engineering Science, Universitas Syiah Kuala.

Since 2001 she has been the lecturer and researcher at the Department of Electrical and Computer Engineering in Universitas Syiah Kuala. Her research interest is developing the e-health monitoring system based on thermal imaging. She is also active in any research related with electrical engineering and deep learning.

Yunida Yunida received her B. Eng degree in Electrical Engineering from Universitas Syiah Kuala, Banda Aceh, Indonesia in 2013. Then she received her Ph.D. in Electrical and Computer Engineering from Universitas Syiah Kuala in 2020 through the "Magister Program of Education Leading to Doctoral for Excellent Graduates (PMDSU)" scholarship from the Ministry of Research, Technology and Higher Education of the Republic of Indonesia. Since 2016, she has published about 6 articles on Scopus Indexed Journals. She is currently a lecturer in the Electrical and Computer Engineering Department at Universitas Syiah Kuala. Her research interests include digital communications, wireless communications, and information theory.