

Lung sound classification using YAMNet, neural network, and augmentation

Jaenal Arifin^{1,3}, Tri Arief Sardjono^{1,2}, Hendra Kusuma¹

¹Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

²Department of Biomedical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

³Electrical Engineering Study Program, Telkom University, Purwokerto, Indonesia

Article Info

Article history:

Received Oct 25, 2024

Revised Jul 16, 2025

Accepted Aug 6, 2025

Keywords:

Augmentation

Classification

Lung sounds

Neural network

YAMNet

ABSTRACT

Globally, lung disease occupies a significant position as one of the main contributors to mortality rates. The characteristics of human respiratory sound signals can show a wide spectrum, ranging from normal patterns to indications of lung abnormalities. The proposed lung sound classification system is based on YAMNet as a pre-trained neural network model for medical audio recognition, which is then refined using artificial neural networks (ANN). This study presents the integration of multiple datasets and advanced pre-processing approaches. A total of 1,363 lung sound recordings from Kaggle, ICBHI, and Mendeley. This reflects the variety of clinical conditions, and differences in recording devices are combined. In order to increase the diversity of lung sound signal input, the pre-processing process is carried out through several stages, including adjusting the sampling frequency to 4 kHz, segmenting for 6 seconds, signal filtering with wavelet, min-max normalization, and data augmentation using window warping, jittering, cropping, and padding. A fold cross-validation scheme is employed to comprehensively evaluate the model's effectiveness. The evaluation results indicate that the model achieves an accuracy of 93.64%, a precision of 93.60%, a recall of 93.64%, and an F1-score of 93.52%, collectively reflecting outstanding classification performance. This work may incorporate deep learning technology into clinical practice, ultimately improving diagnosis accuracy and efficiency in the hospital setting.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Tri Arief Sardjono

Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology

Institut Teknologi Sepuluh Nopember

Keputih Sukolilo, Surabaya, Indonesia

Email: sardjono@bme.its.ac.id

1. INTRODUCTION

Globally, lung disease is a major health issue, and mortality continues to increase every year. Accurate diagnosis of pathological conditions is very important. The sounds produced by the respiratory tract and lungs have important information about the physiological and pathological conditions of humans. Medical professionals often diagnose lung disease by analyzing breath sounds. Movement in the narrowed airways can provide early indications of respiratory distress, such as crackles, wheezes, and rhonchi [1]. The World Health Organization (WHO) projects that chronic obstructive lung disease will be the third greatest cause of mortality by 2030, based on data available to 2022 [2]. A stethoscope is a measuring instrument that can listen to lung sounds and is used by medical personnel [3]. Breath sounds can serve

multiple functions, including consultation, education, and research [4]. Traditional stethoscopes don't record breath sounds for diagnostic purposes. Research on lung sound classification has several challenges, including lung sounds are influenced by various factors such as noise [5], [6], imperfect recording techniques [7], patient physiological conditions, variations in lung sounds between individuals when recording techniques are performed at different times. This condition adds to the complexity and difficulty of classifying lung sounds. A further concern is selecting a deep learning model that aligns with the attributes of the lung sound signal. The selected deep learning model must be able to capture important features of the lung sounds that are the object of research. The development of deep learning technology has provided an alternative solution to the problem of human health examinations. Research on lung sound detection and classification through technology can improve accuracy predictions for lung disease quickly and can be anticipated earlier [8].

Researchers have used artificial intelligence (AI) technology support to classify breathing sounds [9]. The neural network used is deep neural networks (DNN) with the convolution neural networks (CNN) model. The research obtained classification results with an accuracy of 83% for three classes: normal, crackles, and rhonchi. The research proposed by the author aims to classify five categories of lung sounds, namely normal sounds, crepitations (crackles), wheezes, rhonchi, and a combination of crepitations and wheezes. Normal lung sounds can be defined as sounds obtained from the respiratory tract passing through the bronchi and alveoli. This sound is soft and has a low frequency. Crackle lung sounds are abnormal lung sounds, and the characteristics of crackle lung sounds are short and intermittent bursts. These crackles sounds are often associated with pneumonia, indicating the presence of fluid or mucus during the inspiration process. Wheezes lung sounds sound high-pitched or high. Asthma, chronic bronchitis, and chronic obstructive pulmonary disease (COPD) are frequently associated with wheezing and lung sounds [10], [11]. These rhonchi lung sounds have the characteristics of a low and growling sound. This sound is caused by the flow of air in the large airways containing a lot of fluid or mucus. These crackles-wheezes lung sounds are a combination of the two, namely crackles and wheezes. These crackles, wheezes, and lung sounds have the complexity of lung disease involving obstruction and changes that occur in the patient's lung parenchyma [12]. This proposed study uses YAMNet as a feature extractor and an artificial neural network (ANN) to classify five categories, including normal lung sounds, crepitations, wheezes, rhonchi, and a mix of crepitations and wheezes. YAMNet implementations are available in various deep learning frameworks, including TensorFlow [13].

YAMnet is trained using the AudioSet dataset, which allows it to learn lung sound signal patterns. This model has good generalization capacity. This indicates that the model can transform audio signal characteristics, namely lung sound signals, into more precise features. The specific objective of this study is to obtain the best value from the evaluation of the YAMNet model through various metrics such as accuracy, precision, F1-score, and confusion matrix. The real contribution of this study is to modify the pre-processing technique from previous studies [14]. The stages of the pre-processing technique include frequency sampling, segmentation, smoothing wavelet, and min-max normalization. The pre-processing technique used can prepare lung sound data to be more ready and appropriate before being used in the deep learning model. Adding stages of data augmentation techniques from previous studies [15]. The stages of data augmentation include window warping, jittering, cropping, and padding. This augmentation technique has a very important role, especially if you have a limited and unbalanced dataset.

Our novel contribution lies in the integration of multiple lung sound datasets. A total of 1,363 lung sound recordings from Kaggle, ICBHI, and Mendeley. This combination of data encompasses a wide range of clinical conditions and various types of recording devices. Advanced pre-processing involved setting the sampling rate to 4 kHz, segmenting the sounds into 6 second intervals, applying a wavelet transform for signal smoothing, and performing min-max normalization. Data augmentation used window warping, jittering, cropping, and padding. We used a twenty-five-fold cross-validation strategy for model training, optimizing data usage, and ensuring dataset coverage in the classification process. Other contributions, such as embedding from YAMNet, are used as a lung sound feature extractor, and a neural network is used as the main classifier.

2. METHOD

2.1. Database

This study presents the integration of multiple datasets. A total of 1,363 lung sound recordings from Kaggle [16], ICBHI 2017 [17], and Mendeley [18]. The lung dataset collections are openly accessible online for research reasons. The data set from Kaggle was captured using a digital stethoscope connected to a laptop using an amplifier. The lung sound recording data is represented as mono audio data with a bit rate of 705 kbps, a sampling frequency of 44,100 Hz, and a bit depth of 16 bits per sample. The datasets obtained by

Kaggle can be used to identify, categorize, and diagnose lung diseases. The lung sound dataset from the ICBHI 2017 challenge dataset has various health conditions, such as asthma, bronchitis, and pneumonia. The types of lungs sound owned are normal breathing sounds, crackles, wheezes, and a combination of crackles and wheezes. This dataset was initially compiled to facilitate a scientific challenge at the 2017 International Conference on Biomedical Health Informatics (ICBHI). The limitation of the ICBHI 2017 dataset is that it was collected from two different locations with different equipment, so there is a possibility of variability that can affect the results. The effect of using different equipment can affect the quality of the recording. The quality of the resulting recording can vary. With this limitation, it can be a challenge for researchers to process this signal, such as by using a strategy to overcome signal variability, namely by normalization. This can help balance the differences caused by variations in the equipment used during recording.

This dataset has the potential and opportunity to be used as research material, especially with the theme of lung sound classification. The author uses this dataset considering that it can be obtained online, and several researchers have used it for research and publication purposes [19], [20]. The lung sound dataset was acquired at Fortis Hospital, Vasant Kunj, New Delhi, India, by interfacing an electronic stethoscope with a laptop through an amplifier. The amplifier used was designed to amplify lung sound frequencies of 70-2,000 Hz. This range is considered important to ensure that wheezes and crackles can be detected. The amplifier supports frequency and amplifier settings to adjust the frequency and amplifier gain. The lung sound recording process was fed into the laptop via the audio input on the electronic stethoscope. The sampling rate used during the recording was 44,100 Hz, with a single channel with 16 bits per sample and a bit rate of 705 kbps. A total of 1,363 lung sound audio recordings in .wav format were collected from three separate datasets as a database for the experiments in this study. The lung sound datasets utilized in this investigation are summarized in Table 1.

Table 1. Lung sounds multiple datasets

Lung sounds	Sources of data acquisition	Samples
Crackles wheezes	ICBHI and Kaggle	116
Wheezes	Kaggle and ICBHI	232
Crackles	Kaggle and ICBHI	514
Normal	Kaggle	449
Rhonchi	Mendeley data	52
Total		1363

The author uses multiple datasets with the aim of increasing model generalization, larger dataset sizes, and the diversity of clinical conditions. By using datasets from various sources, we will have a wider variety of data. Reducing dataset bias. Using three different datasets can reduce dataset bias and ensure the model is more robust to real conditions. By combining several datasets, it is possible to have more training data. Each dataset in the clinical domain encompasses a range of lung disorders, including asthma, bronchitis, COPD, and pulmonary fibrosis, reflecting their diverse nature. By merging three datasets, the number of samples for each condition increases, enabling the model to more precisely detect and categorize different lung illnesses based on sound patterns.

The stages of this study comprise data acquisition, data preparation, pre-processing, data augmentation, and fold cross-validation, followed by the use of the YAMNet and neural network models. Normal breath sounds, wheezes, crackles, and rhonchi, as well as the coexistence of wheezes and crackles, represent the output classes of the classification process. The application of twenty-five-fold cross-validation is used for model training. The dataset has been partitioned into 25 folds, where in each iteration of the cross-validation process, 24 folds are used for training, while one-fold is allocated for testing.

The primary purpose of using the 25-fold cross-validation scheme in this lung sound classification study is to ensure that the model built can be thoroughly and fairly tested on the entire dataset. The benefit of using 25-fold cross-validation is to maximize data utilization and identify performance imbalances between folds. Accuracy, precision, recall, and F1-score metrics are implemented to assess the model's functionality. The model's capabilities are further enhanced by the visualization of the confusion matrix and the analysis of the training and prediction results. A confusion matrix can be used as a primary indicator in assessing the effectiveness of deep learning models, providing a comprehensive picture of the levels of accuracy, precision, recall, and F1-score achieved. This study proposes an innovative contribution to signal processing, especially in the field of biomedical audio, by utilizing YAMNet as a feature extractor and ANN as a classifier. The proposed investigation adopts a comprehensive methodological framework, in this displayed in Figure 1.

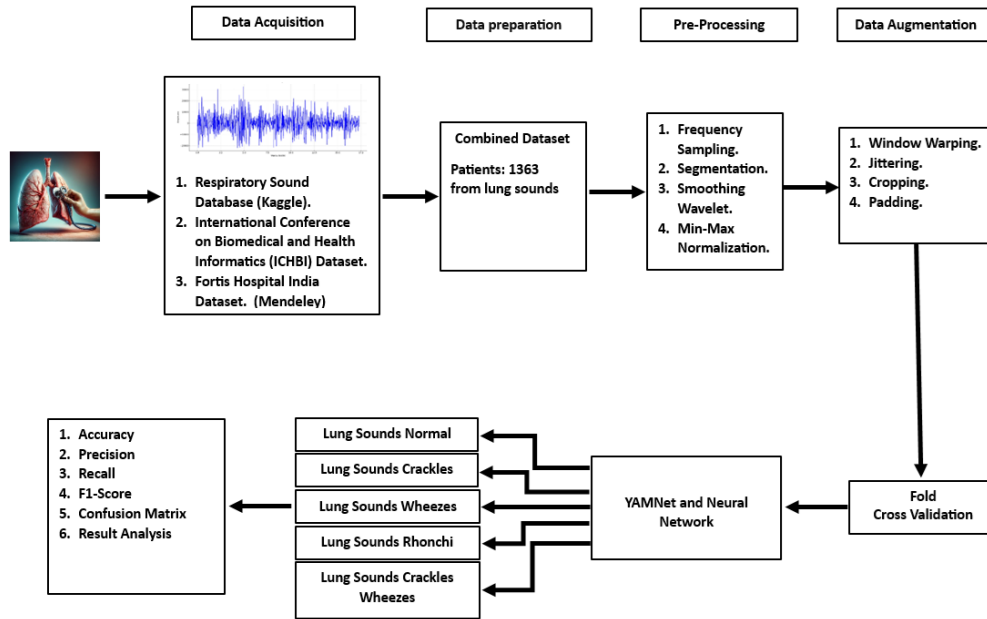


Figure 1. The proposed investigation adopts a comprehensive methodological framework

2.2. Data preparation and pre-processing

The preprocessing workflow comprises four discrete stages, beginning with signal resampling to 4 kHz and segmentation into six second, applying wavelet smoothing, and performing min-max normalization. This text provides comprehensive information about the preprocessing of human lung sound data. Lung sounds predominantly exhibit frequency components in the range of 100 Hz to 2 kHz. According to the Nyquist-Shannon sampling theorem, an analog signal can be reconstructed if the sampling frequency f_s is more than twice the highest frequency of the signal f_{max} [21]. If the highest frequency in a signal is f , then the sampling frequency is at least $2f$. The (1) is by the Nyquist-Shannon sampling theorem.

$$f_s \geq f_{max} \quad (1)$$

Where f_s is sampling frequency and $2f_{max}$ is the maximum frequency present in the signal.

The application of this formula that the sampling frequency (f_s) must be at least twice the maximum frequency (f_{max}) is to avoid alliances. Figure 2 illustrates the sampling of lung sound at 4 kHz in this investigation. The segmentation in this study employed a time interval of 6 seconds. Lung sound segmentation refers to the process of splitting lung sound recordings into distinct and shorter pieces. Segmentation seeks to streamline data management and facilitate analysis. Segmentation is considered essential in deep learning, as it ensures that the input data aligns with the size of the segmentation outcomes, a requirement for deep learning algorithms. Lung sound segmentation of 6 seconds in this study is displayed in Figure 3.

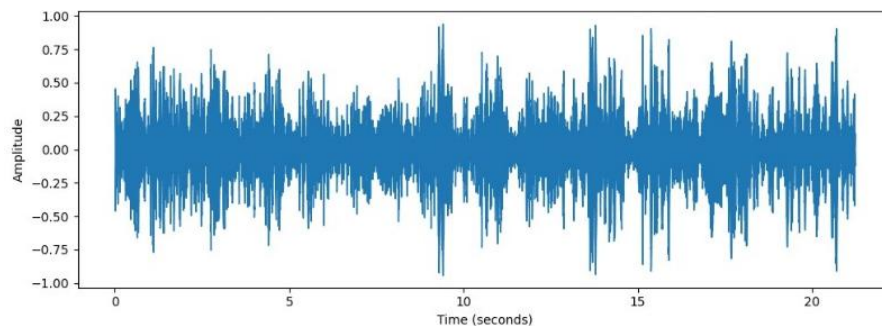


Figure 2. Lung sound frequency sampling 4 kHz

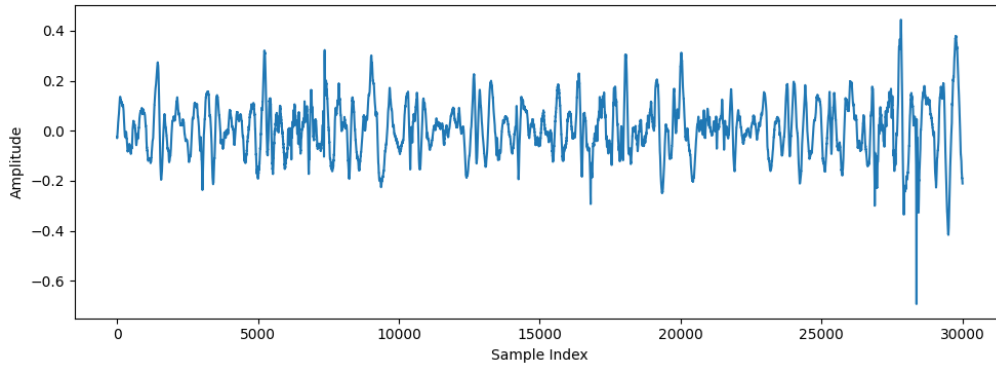


Figure 3. Lung sound segmentation 6 second

Wavelet smoothing is employed in this investigation. Wavelet smoothing is employed to attenuate the noise in the lung sound signal and achieve a smoother signal. There are two main forms of wavelet transform, namely discrete wavelet transform (DWT) and continuous wavelet transform (CWT) [22]. Mathematically, CWT is described as (2).

$$W_f(a, b) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{a}} f(t) \varphi\left(\frac{t-b}{a}\right) dt \quad (2)$$

Where φ is the conjugate complex of the mother wavelet φ , a is the scale factor, and b is the translation factor. CWT is used for detailed signal analysis that requires good accuracy. The process of CWT requires high computing. Mathematically, DWT is described as (3).

$$\text{DWT}[j, k] = \sum_t f(t) \varphi_{j,k}(t) \quad (3)$$

Where DWT $[j, k]$ is the wavelet coefficient at scale j and translation k . $f(t)$ signal function to be transformed. $\varphi_{j,k}(t)$ is wavelet function that depends on the scale j and translation k . The basis wavelet functions $\varphi_{j,k}(t)$ defined as (4).

$$\varphi_{j,k}(t) = \frac{1}{\sqrt{2^j}} \varphi\left(\frac{t-k2^j}{2^j}\right) \quad (4)$$

Where $\varphi(t)$ is mother wavelet. Parameters j and k are used to control the scale and position of the wavelet. DWT can be used to analyze signals. The mother wavelet selection uses the Daubechies-5 (db5) wavelet [23]. The technique employed in this research is min-max normalization. This technique standardizes the signal by ensuring that it has a consistent range of values [24]. The range is defined as a continuum from 0 to 1, which facilitates the process of comparing signals. Min-max normalization is used to mitigate the potential issues that can arise from having a wide range of values [25]. The min-max normalization formula can be given as (5).

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (5)$$

X is the real value, X' is the normalized value, X_{\min} is the minimum value and X_{\max} is the maximum value. One of the advantages of min-max normalization in the context of using deep learning is that it has a fixed range scaling. Fixed scaling is set with a range between 0 and 1. Maintains relationships between data. The values used are after normalization. Using the min-max normalization technique in lung sound signal research is beneficial. The min-max normalization technique can bring all these signals to a uniform range (signal standardization). Min-max normalization normal lung sound in this study is displayed in Figure 4.

2.3. Data augmentation

Improvement of training data can be achieved through the application of data augmentation, especially those with limited and imbalanced datasets. Some reasons that data augmentation techniques can be used are to enlarge the dimensions of the data set, improve model performance, avoid overfitting problems, and fix data imbalance. The subsequent are the data augmentation techniques implemented in this investigation [26].

- i) Window warping: window warping is an augmentation technique that regulates the temporal changes of an audio signal. For example, the temporal changes of a lung sound signal. This change can be in the form of shifting the window and curving it with a dilation or compression procedure [27].
- ii) Jittering: jittering is adding noise data to an audio signal [27]. In this study, Gaussian noise was added to the lung sound signal to add data variation. The noise intensity was randomly selected between 10-20 dB. This technique can help the model to be more robust and help with the variations that occur in real data.
- iii) Cropping: cropping is performed to process this lung sound signal. This technique can be used to focus on a specific part of the lung sound cycle that is relevant for detection and classification [15].
- iv) Padding: the padding function on the lung sound signal is to add a value of 0 (zero) at the end of the signal if its length is less than n-fft (2,048 samples). The padding ensures that further operations, such as cropping, can run smoothly. In this study, the author combined cropping and padding to ensure that all lung signals have a uniform length to be used as model input. The authors refer to research that incorporates these augmentation techniques in the context of image data, and similar principles can be applied to audio data [28].

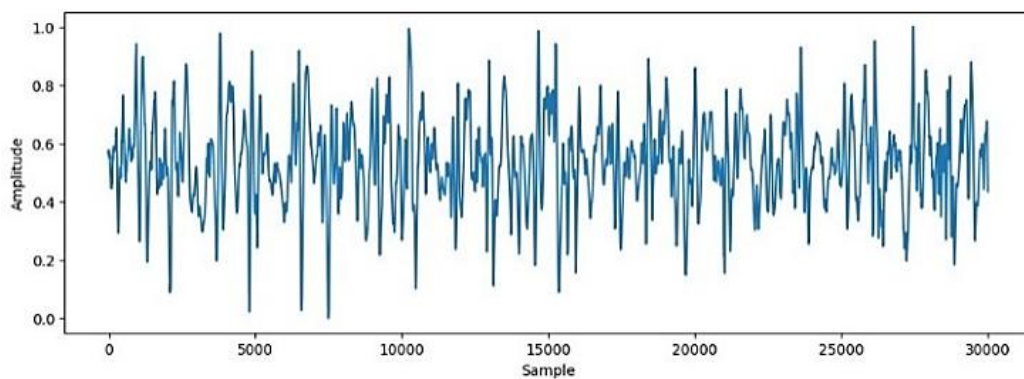


Figure 4. Min-max normalization normal lung sound

2.4. Fold cross validation

This technique can decrease variability, enhance data use, and achieve a balance between bias and variation [29], [30]. Cross-validation was carried out by the authors, with configurations scaling up to 25 folds. The 25-fold cross-validation approach involves dividing the dataset into 25 approximately balanced portions, or folds, based on the number and distribution of samples. Each component exhibits a representative distribution of the complete dataset. The model undergoes 25 iterations of training and testing. In each iteration, one of the 25 folds is chosen as the test data, while the remaining 24 folds are utilized for training the model. This guarantees that each fold is utilized once as the test data and 24 times as part of the training data. The purpose of performing cross-validation up to 25-fold is to assess how well the deep learning model performs in the training process and evaluate the model's performance. After the iteration is complete, the evaluation results from 25 folds are collected to obtain metric values. Confusion matrix in this study is displayed in Table 2.

Table 2. Confusion matrix

Actual/Predicted	Normal	Crackles	Wheezes	Rhonchi	Crackles	Wheezes
Normal	TP	FP	FP	FP	FP	FP
Crackles	FN	TP	FP	FP	FP	FP
Wheezes	FN	FN	TP	FP	FP	FP
Rhonchi	FN	FN	FN	TP	FP	FP
Crackles Wheezes	FN	FN	FN	FN	FP	TP

2.5. Proposed YAMNet and neural network

The YAMNet model can extract various attributes from audio recordings, including frequency, amplitude, and sound texture. These features are relevant for audio signal identification and classification. Embeddings from YAMNet are used as feature representations of lung sounds. These embeddings are vectors

with fixed dimensions that describe the characteristics of the lung sounds. After the features are extracted, the neural network model performs classification based on these features. This neural network model (fully connected) has several dense layers (2048, 1024, 512, 256, 128 neurons) and each layer incorporates a rectified linear unit (ReLU) activation, followed by batch normalization and dropout operations to promote stable convergence and minimize overfitting. The SoftMax function is utilized in the output layer to produce a probabilistic distribution over the predefined categories of lung sounds, namely normal, wheezes, crackles, rhonchi, and the combination of crackles and wheezes. As shown in Figure 5, the workflow outlines the structural design of the model developed in this study.

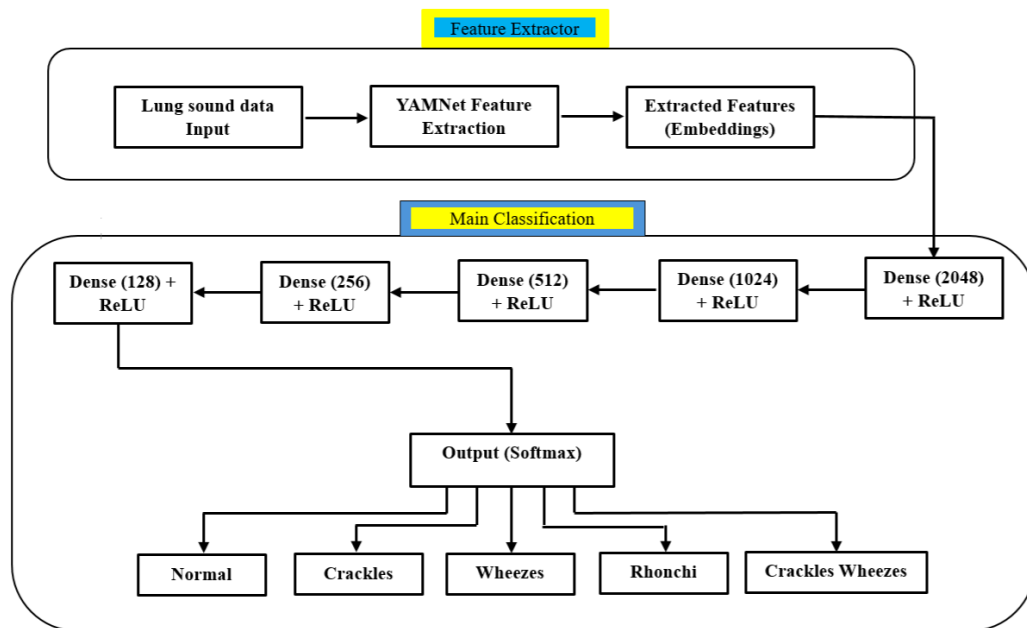


Figure 5. Works flow a model

3. RESULTS AND DISCUSSION

K-fold cross-validation experiments were conducted. As part of the cross-fold validation process, we reviewed the effects of different cross-fold values on our lung-para sound classification model. The authors performed tests with settings from 1 to 25. The authors found that the three best accuracies were 97.24% epoch 16, 97.06% epoch 18, and 96.88% epoch 17. Accuracy of different folds in this study is displayed in Figure 6.

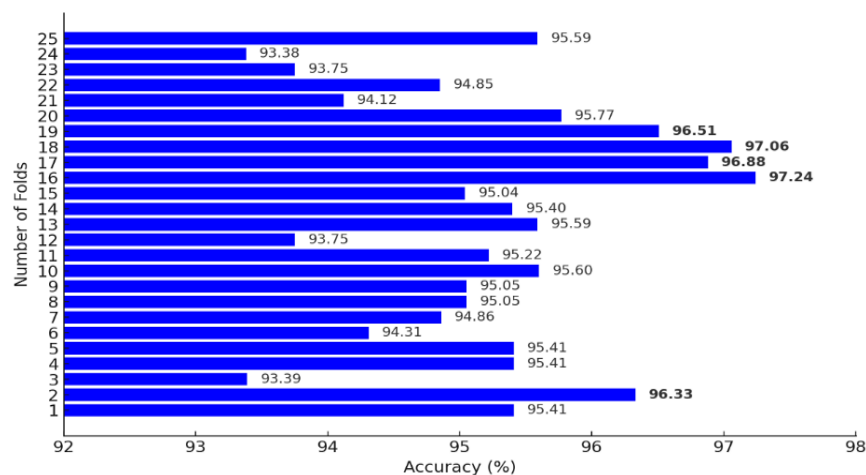


Figure 6. Accuracy of different folds

The authors evaluated the performance of various optimizers as objects in a lung sound classification model and recorded their accuracy values as part of the performance analysis. The AdaMax optimizer shows the best performance in all metrics, with an accuracy of 93.42%, precision of 93.38%, recall of 93.42%, and F1-score of 93.29%. Optimizer results in this study are displayed in Table 3.

The AdaMax optimizer demonstrates a well-balanced performance across recall, precision, and F1-score metrics. This very good balance indicates that this model works well, namely the ability to detect positive classes correctly (precision) and the ability to detect all positive classes (recall). The second best after AdaMax is the Adam optimizer, with an accuracy of 91.62%, precision of 91.60%, recall of 91.62%, and F1-score of 91.40%. The precision and recall values are consistent, indicating good performance in detecting positive and negative classes. Balanced performance is evidenced by the comparable values of recall, precision, and F1-score, indicating that there is no significant bias in any of the metrics. The performance of the Adam optimizer is very good; it can be the second choice after the AdaMax optimizer. The third best optimizer sequence is stochastic gradient descent (SGD), with an accuracy of 89.30%, balanced precision, and recall values. SGD is a choice if the model requires a simpler learning process, although it has slightly lower performance than the Adam and AdaMax optimizers. The fourth best optimizer is RMSprop, with an accuracy of 87.72%, lower than the previous three optimizers. The precision and recall values are balanced, and the F1-score value is slightly lower (87.22%). The RMSprop optimizer provides quite good values, but its performance is lower than that of Adam, AdaMax, and SGD. The adaptive gradient (Adagrad) optimizer is not recommended for this lung sound classification task because it has a low accuracy of 46.58%. This indicates that the model is not able to learn well and is unable to detect positive and negative classes.

The authors investigate the impact of changing the number of epochs on the effectiveness of our model for lung sound classification. Our starting point is 10 epochs; there is a significant increase in metrics when the number of epochs increases from 10 to 100 epochs. This performance is relatively stable. At epoch 50, there is a significant increase in values in all metrics, with accuracy increasing to 84.42%. At epoch 100, all metrics are around 91.55 to 91.59%. This shows that the model has learned very well. At epoch 150 to 300, the accuracy and other metrics reach 93%. This model's optimal epoch is 200 to 300; it achieves the best performance with high and stable accuracy, precision, recall, and F1-score values. The best accuracy is at 93.64%; this occurs at epoch 300. Epochs table in this study is displayed in Table 4.

After epoch 300, there is a slight decrease in performance in several metrics, such as F1-score and recall. This indicates that increasing the number of epochs can no longer improve performance on the testing data. At epoch 350 and 400, performance stability with minor declines occurred. The accuracy, precision, recall, and F1-score values decreased slightly to 93.20%, indicating that the model began to have difficulty maintaining its performance.

Table 3. Evaluation metrics of different optimizers used in the proposed model

Optimizer	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Adam	91.62	91.60	91.62	91.40
RMSprop	87.72	87.73	87.73	87.22
Adagrad	46.58	46.15	46.83	44.14
SGD	89.30	89.34	89.30	88.85
AdaMax	93.42	93.38	93.42	93.29

Table 4. Model performance across different epochs

Number of epochs	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
10	59.95	63.52	59.95	59.17
50	84.42	84.05	84.42	84.08
100	91.55	91.59	91.55	91.12
150	92.76	92.87	92.76	92.56
200	93.42	93.40	93.42	93.29
250	93.24	93.18	93.24	93.08
300	93.64	93.60	93.64	93.52
350	93.20	93.18	93.20	93.05
400	93.20	93.14	93.20	93.06

The confusion matrix with the Adamax optimizer shows that YAMNet and the neural network models have excellent performance. The models were able to classify the rhonchi and crackles_wheezes classes perfectly. The wheezes class was also classified accurately. The model was able to recognize the typical characteristics of the wheezes class. The normal and crackle classes need to be improved to distinguish between them. This study presents the confusion matrix for lung sound categorization as the object, which is displayed in Figure 7 at epoch 400.

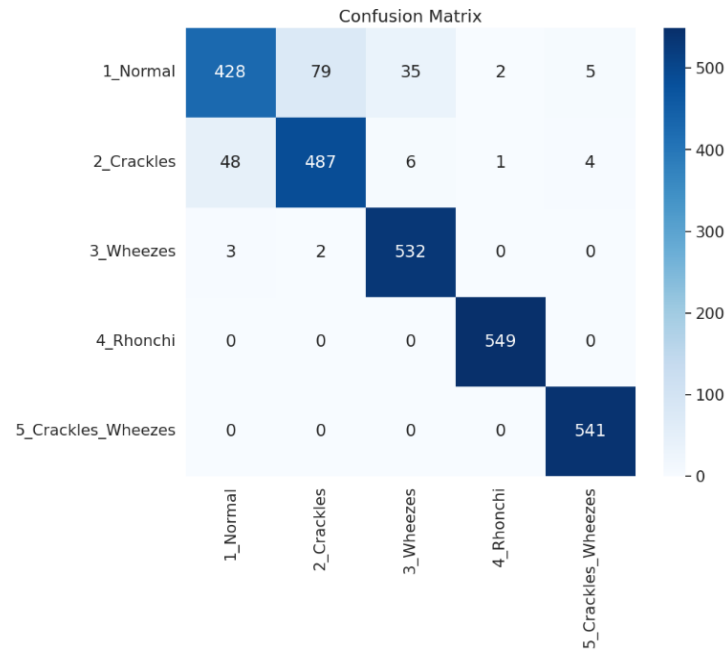


Figure 7. Confusion matrix for lung sound categorization

The study includes a performance comparison of the proposed models with previously established approaches. There was an increase in the comparison results obtained by the author from 2022 to 2025. A comparative evaluation between the proposed method and existing methodologies displayed in Table 5. Based on the comparison of the Kaggle dataset, the neural network model got an F1-score of 72.41%.

The YAMNet and neural network models got a score of 91.4%. Based on the comparison using only the ICBHI dataset, the proposed model (YAMNet, neural network) got the highest performance with an F1-score of 98.52%. The CNN+bidirectional long short-term memory (BDLSTM) model recorded an F1-score of 95.90%. The short-time Fourier transform (STFT)-CNN model F1-score is lower at 78.45%. More modern models, such as residual attention network-based vision transformer (RAN-ViT), achieve an F1-score of 97.49%. The multi-task learning model F1-score of 98.40%. Proposed methods such as models (YAMNet+neural network) and multi-task learning show promising research directions in the field of medical sound classification. Based on the combined ICBHI and Mendeley dataset, the proposed model (YAMNet+neural network) achieved an F1-score of 99.63%. This value is higher than the lightweight CNN F1-score model of 97.80%. The addition of datasets (ICBHI, Kaggle, and Mendeley) does not guarantee better performance. Performance drops to an F1-score of 93.64%. A comparison of accuracy based on the dataset shows that the ICBHI+Mendeley dataset produces the highest accuracy of 98.67%. Adding datasets (ICBHI+Kaggle+Mendeley) can reduce accuracy (93.06%). The Kaggle dataset produces the lowest accuracy (82.26%) compared to others. This highlights the limitations of the data quality if used alone.

Table 5. A comparative evaluation between the proposed method and existing methodologies

Method	Dataset	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Neural network [31], (2022)	Kaggle	73.67	72.96	72.41	72.96
CNN + BDLSTM [14], (2022)	ICBHI,	95.91	95.90	95.90	95.90
	King Abdullah University Hospital				
STFT-CNN [22], (2023)	ICBHI	79.45	78.09	78.45	78.09
Multi-feature integration utilizing lightweight CNN model [32], (2024)	ICBHI, Mendeley	97.76	97.70	97.70	97.70
Combined RAN-ViT [33], (2024)	ICBHI	97.54	97.49	97.49	97.49
YAMNet [34], (2024)	ICBHI, Kaggle	97.94	97.92	97.92	97.92
Multi-task learning [35], (2025)	ICBHI	98.42	98.40	98.40	98.40
Proposed method	ICBHI	98.56	98.53	98.52	98.53
(YAMNet, neural network)	Kaggle	91.39	91.55	91.4	91.55
	ICBHI, Mendeley	99.64	99.63	99.63	99.63
	ICBHI, Kaggle, Mendeley	93.64	93.64	93.64	93.06

4. CONCLUSION

The integration of YAMNet for feature extraction and a neural network as the primary classification model presents promising prospects for analyzing lung sounds, as well as other applications in the biomedical field. With proper adjustment and training, YAMNet and neural networks can be very useful tools for improving diagnosis and health monitoring in the medical field. Testing of various optimization algorithms to see their effect on model accuracy. The AdaMax optimizer provides the best accuracy value of 93.42%, precision 93.38%, recall 93.42, and F1-score 93.29%, compared to other optimizers (Adam, RMSprop, Adagrad, and SGD), which produce lower metric values (based on result and discussion). The optimal epoch for this model recorded an accuracy of 93.64%, a precision of 93.60%, a recall of 93.64%, and an F1-score of 93.52%; this occurred at epoch 300.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the providers of the Kaggle, ICBHI 2017 Challenge, and Mendeley Data repositories for making their datasets openly available, thereby supporting advancements in this work.

FUNDING INFORMATION

This work was supported by the Government of Indonesia, Center for Higher Education Funding and Assessment (PPAPT), in part by the Indonesian Endowment Fund for Education (LPDP), and in part by the Indonesian Education Scholarship (BPI), under Grant 202231103922.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Jaenal Arifin	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Tri Arief Sardjono	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓		
Hendra Kusuma	✓	✓	✓	✓		✓	✓			✓	✓	✓	✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest related to the research and publication of this paper.

DATA AVAILABILITY

The data used in this study are publicly available from the following sources:





- Kaggle, open available at <https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database>, reference number [16].
- ICBHI 2017, open accessible via https://bhichallenge.med.auth.gr/ICBHI_2017_Challenge, reference number [17].
- Mendeley Data, open available at <https://data.mendeley.com/datasets/fr7zvy8j5s/1>, reference number [18].

REFERENCES





- [1] A. T. Abdulahi, R. O. Ogundokun, A. R. Adenike, M. A. Shah, and Y. K. Ahmed, "PulmoNet: a novel deep learning based pulmonary diseases detection model," *BMC Medical Imaging*, vol. 24, no. 1, 2024, doi: 10.1186/s12880-024-01227-2.
- [2] WHO, "The top 10 causes of death," *World Health Organization*. 2024. Accessed: Aug. 22, 2024 [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>
- [3] B. A. Tessema, H. Nemomssa, and G. L. Simegn, "Acquisition and classification of lung sounds for improving the efficacy of auscultation diagnosis of pulmonary diseases," *Medical Devices: Evidence and Research*, vol. 15, pp. 89–102, 2022, doi: 10.2147/MDER.S362407.

- [4] D.-M. Huang, J. Huang, K. Qiao, N.-S. Zhong, H.-Z. Lu, and W.-J. Wang, "Deep learning-based lung sound analysis for intelligent stethoscope," *Military Medical Research*, vol. 10, no. 1, 2023, doi: 10.1186/s40779-023-00479-3.
- [5] G. Petmezaz *et al.*, "Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function," *Sensors*, vol. 22, no. 3, 2022, doi: 10.3390/s22031232.
- [6] H. Liu *et al.*, "MEMS piezoelectric resonant microphone array for lung sound classification," *Journal of Micromechanics and Microengineering*, vol. 33, no. 4, 2023, doi: 10.1088/1361-6439/acbf3.
- [7] T. Nguyen and F. Pernkopf, "Lung sound classification using co-tuning and stochastic normalization," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 9, pp. 2872–2882, 2022, doi: 10.1109/TBME.2022.3156293.
- [8] Y. Kim *et al.*, "Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning," *Scientific Reports*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-96724-7.
- [9] F. Cinyol, U. Baysal, D. Köksal, E. Babaoğlu, and S. S. Ulaşlı, "Incorporating support vector machine to the classification of respiratory sounds by convolutional neural network," *Biomedical Signal Processing and Control*, vol. 79, 2023, doi: 10.1016/j.bspc.2022.104093.
- [10] R. Palaniappan, K. Sundaraj, and N. U. Ahamed, "Machine learning in lung sound analysis: a systematic review," *Biocybernetics and Biomedical Engineering*, vol. 33, no. 3, pp. 129–135, 2013, doi: 10.1016/j.bbe.2013.07.001.
- [11] A. H. Sfayyih, N. Sulaiman, and A. H. Sabry, "A review on lung disease recognition by acoustic signal analysis with deep learning networks," *Journal of Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00762-z.
- [12] A. Agustí *et al.*, "Global initiative for chronic obstructive lung disease 2023 report: GOLD executive summary," *European Respiratory Journal*, vol. 61, no. 4, 2023, doi: 10.1183/13993003.00239-2023.
- [13] Google, "YamNet: an audio event classifier trained on the AudioSet dataset to predict audio events from the AudioSet ontology," *Kaggle*. 2020. Accessed: Aug. 22, 2024. [Online]. Available: <https://www.kaggle.com/models/google/yamnet>
- [14] M. Fraiwan, L. Fraiwan, M. Alkhodari, and O. Hassanin, "Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 10, pp. 4759–4771, 2022, doi: 10.1007/s12652-021-03184-y.
- [15] J. A. Dar, K. K. Srivastava, and A. Mishra, "Lung anomaly detection from respiratory sound database (sound signals)," *Computers in Biology and Medicine*, vol. 164, 2023, doi: 10.1016/j.compbiomed.2023.107311.
- [16] Vbookshelf, "Respiratory sound database," *Kaggle*. 2018. [Online]. Available: <https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database>
- [17] B. M. Rocha *et al.*, "ICBHI 2017 challenge: respiratory sound database," *ICBHI Challenge*, 2019. https://bhichallenge.med.auth.gr/ICBHI_2017_Challenge
- [18] M. K. Dutta, "Pulmonary (lungs) sound," *Mendeley Data*, ver. 1, 2022, doi: 10.17632/fr7zvy8j5s.1.
- [19] T. Nguyen and F. Pernkopf, "Crackle detection in lung sounds using transfer learning and multi-input convolutional neural networks," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, 2021, pp. 80–83, doi: 10.1109/EMBC46164.2021.9630577.
- [20] A. F. R. Gomez and A. D. O.-Canon, "Respiratory sounds classification employing a multi-label approach," in *2021 IEEE Colombian Conference on Applications of Computational Intelligence*, 2021, pp. 1–5, doi: 10.1109/ColCACI52978.2021.9469042.
- [21] H. Johansson, "Sampling and quantization," in *Signal Processing and Machine Learning Theory*, Cambridge, Massachusetts: Academic Press, 2024, pp. 185–265, doi: 10.1016/B978-0-32-391772-8.00011-9.
- [22] O. Gazi, *Understanding digital signal processing*. Singapore: Springer, 2018, doi: 10.1007/978-981-10-4962-0.
- [23] N. K. Kularathne, M. M. P. M. Fernando, and J. M. U. T. D. Jayasinghe, "High-frequency noise removal of audio files using daubechies wavelet transform," *Current Scientia*, vol. 26, no. 2, pp. 57–63, 2023, doi: 10.31357/vjs.v26i02.6807.
- [24] M. Shantal, Z. Othman, and A. A. Bakar, "A novel approach for data feature weighting using correlation coefficients and min-max normalization," *Symmetry*, vol. 15, no. 12, 2023, doi: 10.3390/sym15122185.
- [25] S. A. D. Prasetyowati, M. Ismail, E. N. Budisusila, D. R. I. M. Setiadi, and M. H. Purnomo, "Dataset feasibility analysis method based on enhanced adaptive LMS method with min-max normalization and fuzzy intuitive sets," *International Journal on Electrical Engineering and Informatics*, vol. 14, no. 1, pp. 55–75, 2022, doi: 10.15676/ijeei.2022.14.1.4.
- [26] K. N. Lal, "A lung sound recognition model to diagnoses the respiratory diseases by using transfer learning," *Multimedia Tools and Applications*, vol. 82, no. 23, pp. 36615–36631, 2023, doi: 10.1007/s11042-023-14727-0.
- [27] Q. Pan, X. Li, and L. Fang, "Data augmentation for deep learning-based ECG analysis," in *Feature Engineering and Computational Intelligence in ECG Monitoring*, Gateway East, Singapore: Springer Publishing, Jun. 2020, pp. 91–111, doi: 10.1007/978-981-15-3824-7_6.
- [28] R. Takahashi, T. Matsubara, and K. Uehara, "RICAP: random image cropping and patching data augmentation for deep CNNs," *Proceedings of Machine Learning Research*, vol. 95, pp. 786–798, 2018.
- [29] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: a review with examples from ecology," *Ecological Monographs*, vol. 93, no. 1, 2023, doi: 10.1002/ecm.1557.
- [30] J. Kaliappan, A. R. Bagepalli, S. Almal, R. Mishra, Y.-C. Hu, and K. Srinivasan, "Impact of cross-validation on machine learning models for early detection of intrauterine fetal demise," *Diagnostics*, vol. 13, no. 10, pp. 1–22, May. 2023, doi: 10.3390/diagnostics13101692.
- [31] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "A neural network-based method for respiratory sound analysis and lung disease detection," *Applied Sciences*, vol. 12, no. 8, 2022, doi: 10.3390/app12083877.
- [32] T. Wanasinghe, S. Bandara, S. Madusanka, D. Meedeniya, M. Bandara, and I. D. L. T. Díez, "Lung sound classification with multi-feature integration utilizing lightweight CNN model," *IEEE Access*, vol. 12, pp. 21262–21276, 2024, doi: 10.1109/ACCESS.2024.3361943.
- [33] M. Jurej, R. Roslidar, and Y. Yunida, "Improved lung sound classification model using combined residual attention network and vision transformer for limited dataset," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 12, no. 4, pp. 1085–1097, 2024, doi: 10.52549/ijeei.v12i4.5530.
- [34] H. Widyagustin, H. Kusuma, and T. A. Sardjono, "Deep learning-based classification of lung sound using YAMNet for identifying lung diseases," in *2024 2nd International Symposium on Information Technology and Digital Innovation*, 2024, pp. 131–135, doi: 10.1109/ISITDI62380.2024.10796009.
- [35] K. V. Suma, D. Koppad, P. Kumar, N. A. Kantikar, and S. Ramesh, "Multi-task learning for lung sound and lung disease classification," *SN Computer Science*, vol. 6, no. 1, 2024, doi: 10.1007/s42979-024-03506-9.





BIOGRAPHIES OF AUTHORS

Jaenal Arifin     received the bachelor degree in Electrical Engineering at Diponegoro University in 2004 and a master's degree in Electrical Engineering at Gadjah Mada University in 2016. He is currently pursuing a doctoral degree in the Department of Electrical Engineering at Institut Teknologi Sepuluh Nopember. He is a lecturer at Telkom University Purwokerto. His research interests include electronics, signal processing, and biomedical electronics. He can be contacted at email: 7022221021@student.its.ac.id or jaetoga33@gmail.com.



Tri Arief Sardjono     received the bachelor degree in Electrical Engineering at Institut Teknologi Sepuluh Nopember and the master degree in Biomedical Engineering Program from Institut Teknologi Bandung, in 1994 and 1999, respectively, and the Ph.D. degree from the University of Groningen, in 2007. Since 1995, he has been working with the Department of Electrical Engineering and Biomedical Engineering, Institut Teknologi Sepuluh Nopember. His research interest includes biomedical image processing and analysis. He can be contacted at email: sardjono@bme.its.ac.id.



Hendra Kusuma     received the B.S. and Ph.D. degree from Institut Teknologi Sepuluh Nopember Surabaya both in Electrical Engineering, in 1988 and 2016, respectively. He also received the M.S. degree from Curtin University of Technology in Renewable Energy Engineering in 2001. From 1989 until now, he is a lecturer in the Department of electrical engineering, Institut Teknologi Sepuluh Nopember Surabaya. His research interests are in artificial intelligence, machine learning, assistive technology, and IoT as well as in applied electronic. He can be contacted at email: hendraks@ee.its.ac.id.