Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 12, No. 1, March 2024, pp. 101~112
ISSN: 2089-3272, DOI: 10.52549/ijeei.v12i1.5343

Classification of Human Emotions Based on Javanese Speech Using Convolutional Neural Network and Multilayer Perceptron

Muji Ernawati[1], Dwiza Riana[2]
[1,2]Computer Science, Department of Information Technology, Universitas Nusa Mandiri, Indonesia

Article Info

Article history:

Received Dec 23, 2023
Revised Feb 25, 2024
Accepted Mar 10, 2024

Keywords:

Speech Emotion Recognition
Convolutional Neural Network
Multilayer Perceptron
Data Augmentation
Mel Frequency Cepstral Coefficients

ABSTRACT

Emotions in speech are considered a basic principle of human interaction and play an important role in decision-making, learning, and everyday communication. Research on speech emotion recognition is still being carried out by many researchers to develop speech emotion recognition models with better performance. In this research, we combine the application of data augmentation techniques (Add Noise, Time Stretch, and Pitch Shift) to increase the data size of the Javanese Speech Emotion Database (Java-SED). Mel Frequency Cepstral Coefficients (MFCC) is used for feature extraction and then builds a Convolutional Neural Network (CNN) model and applies Multilayer Perceptron (MLP) to classify human emotions from sound. In this research, we produced eight experimental models with a combination of different augmentation techniques. The CNN model parameters include 40 input neurons, four hidden layers with varying neuron counts, Relu activation functions, L2 regularization, dropout rates, Adam optimization, and ModelCheckpoint callbacks to save the best model based on validation loss. From the results of the evaluation that has been carried out, the CNN algorithm produces the highest performance with an accuracy of 96.43%, recall of 96.43%, precision of 96.57%, F1-score of 96.48%, and kappa of 95.71% by applying the Add Noise technique, Time Stretch, and Pitch Shift.

Corresponding Author:

Dwiza Riana,
Department of Information Technology,
Universitas Nusa Mandiri,
Jl. Jatiwaringin Raya No.2, Makasar, Jakarta Timur 13620, DKI Jakarta, Indonesia.
Email: 14210225@nusamandiri.ac.id

1. INTRODUCTION

Emotions are a form of expression for humans that will naturally appear in everyday conversations to express their feelings clearly [1]. Emotions in speech are considered a basic principle of human interaction and play an important role in decision-making, learning, and everyday communication [2]. Much research has been conducted to understand human emotions from sounds emitted by machines, for example, studies on personal digital assistants, call-center systems, speech-to-text models and sensor applications [3]. This has led to the recognition of human emotions through sound being considered one of the most important research areas in the last decade.

Deep learning is one of the most widely used speech emotion recognition modeling architectures because its main advantage lies in the automatic selection of features. For example, it can be applied to important attributes attached to sound files that have a specific task for recognizing speech emotions [4]. Multilayer Perceptron (MLP) is a powerful deep learning algorithm with the ability to classify very complex information [5]. The advantage of MLP lies in determining better weight values compared to other methods [6]. Researchers also often use MLP to detect human emotions from sound [5], [7]–[9]. Apart from that, the

Convolutional Neural Network (CNN) is a deep learning algorithm extensively applied in the domain of computer vision and also in research on human emotions based on sound, such as [4], [10]–[12]. CNN stands out with its strong capability to comprehend patterns from a majority of samples and represent knowledge at a higher level, while providing an advantage in reducing the number of parameters during training through weight sharing [13]. The simultaneous feature extraction process carried out by this network makes it a highly organized system and easier to implement compared to other types of networks. The CNN algorithm is more efficient than other neural network algorithms, especially in terms of memory and complexity. The need for large data sets is one of the problems with deep learning algorithms, especially CNN, which is a data-hungry model [14].

Data augmentation is a method for increasing the amount of data by adding copies that have been slightly modified or synthetic data that is recreated from existing samples [15]. When the amount of data owned is small or not considered sufficient to improve model performance, applying data augmentation techniques can be a solution to add new data samples. As in previous studies, the use of data augmentation methods like time stretching, noise addition, and pitch shift allows new data samples to be generated and model performance to be improved [10], [16], [17].

This research aims to apply several data augmentation techniques (add noise, time stretch, and pitch shift) to increase the data size and is expected to improve model performance using a public dataset, namely the Javanese Speech Emotion Database (Java-SED) [18]. Then, apply Mel-Frequency Cepstral Coefficients (MFCC) feature extraction to help obtain features and characteristics from voice data. And build a CNN architecture model and use the MLP model from previous research [18] to classify human emotions based on sound. This research article consists of an explanation of the research method in the first section. It then describes the dataset used, the augmentation techniques employed, the feature extraction techniques utilized, a description of the proposed classification model, and the performance metrics employed. Following that, the results of the conducted experiments and a discussion of the achieved results are provided. Finally, conclusions are drawn from the findings of this research.

2. RESEARCH METHOD

Classifying human emotions based on sound goes through research stages, as shown in Figure 1. The first step is analyzing the dataset, then dividing the training and testing data. Before performing feature extraction, data augmentation is applied to the training data. Then classify human emotions using CNN and MLP algorithms.

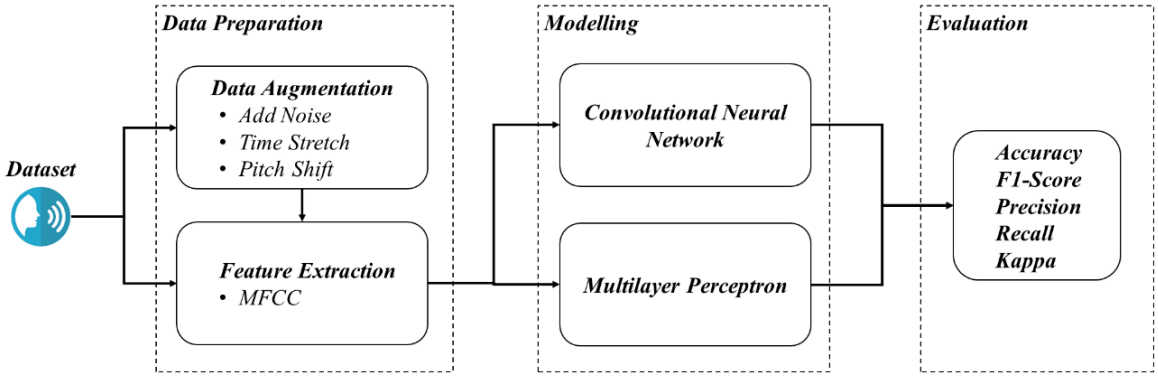

Figure 1. Stages of research

In this study, the Javanese Speech Emotion Database (Java-SED) is utilized, encompassing six emotional classes in the Javanese "Ngoko" language and involving ten experienced actors from the Kamasetra community. The dataset consists of 1680 audio data in WAV format. In the data preparation stage, the dataset is divided into training and testing data, with 80% of the training data used to train the model. Subsequently, data augmentation is performed on the training data using techniques like Add Noise, Time Stretch, and Pitch Shift. The MFCC feature extraction method is then applied to the augmented data to detect human emotions, producing 40 features representing audio data. This process serves as the foundation for further analysis related to emotion recognition. In the modeling phase, an experimental approach is employed by combining three data augmentation techniques to evaluate the impact of adding new data to the training set on improving the performance of the classification model. The resulting eight experimental models involve the application of the MFCC feature extraction technique and the use of CNN and MLP algorithms for human emotion classification. The parameters of the CNN model consist of 40 input neurons,

four hidden layers with different numbers of neurons, Relu activation functions, L2 regularization, dropout rates, Adam optimization, and ModelCheckpoint callbacks for saving the best model determined by validation loss. This experiment aims to evaluate the effectiveness of different augmentation strategies in emotion detection, forming the basis for further analysis and model optimization. A variety of performance indicators, such as accuracy, recall, precision, f1-score, and kappa, are used to evaluate the model. This evaluation provides a profound understanding of the model's capability in classifying human emotions, considering crucial aspects such as accuracy, precision, and recall.

2.1. Dataset

The dataset used in this research is the Javanese Speech Emotion Database (Java-SED) [18]. This dataset is a dataset that uses one of the regional languages in Indonesia, namely the Javanese language "Ngoko". There are 6 emotional classes in this dataset, namely happiness, neutral, sadness, anger, fear, and surprise. In this dataset, there are 10 actors aged between 20 and 30 years, with five female actors and five male actors. The speakers were experienced practitioners who are members of the Kamasetra community at Yogyakarta State University, Indonesia. Each speaker conveyed six different emotions. Each emotion has four types of sentences. Each sentence in one emotion is repeated seven times. Thus, 168 recorded data were obtained from each speaker. The database contains a total of 1680 audio data in WAV format, where each emotion class has 280 data.

2.2. Data Preparation

In the initial phase of preprocessing, the dataset is divided through a process of data splitting into two segments: training data and testing data. 80% of the training data is used to train the model being built, then 20% of the validation data and testing data are used to test the model.

Then, in several experimental models the training data was augmented, with data augmentation techniques added to produce new synthetic data. Add Noise is a method for adding white noise [10]. White noise is a random value inserted into audio data. The acceptable noise amplitude range is $\sigma \in [0.001,0.015]$ [19]. Time Stretch is an audio data augmentation technique that is used to change the speed or length of an sound flag without influencing its pitch [10]. Time Stretch changes the rhythm and length of an sound clip with a stretch calculate range $\gamma \in [0.8,1.25]$ [19]. Pitch Shift is the method of changing the pitch without influencing the speed [10]. Pitch Shift increases or decreases the pitch of an audio sample (while leaving its duration unchanged) by n_ steps semitones. The range of n_steps that can be used for augmentation is $\in$[-4,4] [19].

The next stage is extracting data by presenting the Feature Extraction method on MFCC from data that has gone through an augmentation process to detect human emotions. MFCC offers a brief representation of the assessed votes for ensuing investigation. The Mel scale is derived from a nonlinear frequency scale transformation designed to closely approximate the human perceptual ability to detect small changes in pitch at both low and high frequencies [20]. The stage process of MFCC can be seen in Figure 2. The mfcc feature extraction process in this research produces 40 features that represent audio data.

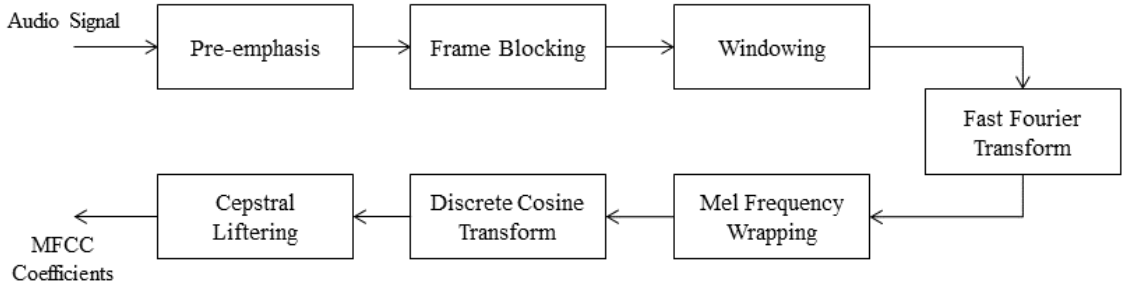

Figure 2. MFCC Process Stages

2.2. Modelling

The experimental model was carried out by combining three data augmentation techniques to see the effect of adding new data to the training data on improving the performance of the classification model. Table 1 is the experimental model that will be carried out in this research based on a combination of the Add Noise, Time Stretch, and Pitch Shift techniques.

Table 1. Experimental Model

Model	Data Augmentation	Feature Extraction
A	Training Data	MFCC
B	Training Data + Add Noise	MFCC
C	Training Data + Time Stretch	MFCC
D	Training Data + Pitch Shift	MFCC
E	Training Data + Add Noise + Time Stretch	MFCC
F	Training Data + Add Noise + Pitch Shift	MFCC
G	Training Data + Time Stretch + Pitch Shift	MFCC
H	Training Data + Add Noise + Time Stretch + Pitch Shift	MFCC

Experiments were carried out on eight models, each applying the MFCC feature extraction technique and modeling human emotion classification using the CNN and MLP algorithms. The MLP architecture in this research uses architecture from previous research [18]. Meanwhile, the CNN architecture used in this research is presented in Figure 3. In the proposed architecture, this study utilizes a one-dimensional Convolutional Neural Network (Conv1D) to classify human emotions based on features extracted from audio files. The first Conv1D layer receives an input array of size 40 x 1, consisting of 32 filters with a kernel size of 3, strides of 1, and kernel regularization of 0.0001. The output is activated by Rectifier Linear Units (ReLu), followed by batch normalization and a dropout of 0.07. The subsequent Conv1D layer employs 64 filters with the same kernel and strides, processing the output of the previous layer with ReLU, batch normalization, and a dropout of 0.07, and then feeds the output into a max-pooling1D layer with a pool size of 2 and strides of 1. This process is repeated in the next two convolutional layers, each with dropouts of 0.07 and 0.14, respectively. The output from the flatten layer is then directed to a fully connected layer with 6 units, corresponding to the predicted number of classes. Softmax activation is applied, followed by regularization of 0.0001. The proposed model uses the Adam optimizer with a learning rate set at 0.0001 and employs categorical cross-entropy as the loss function.

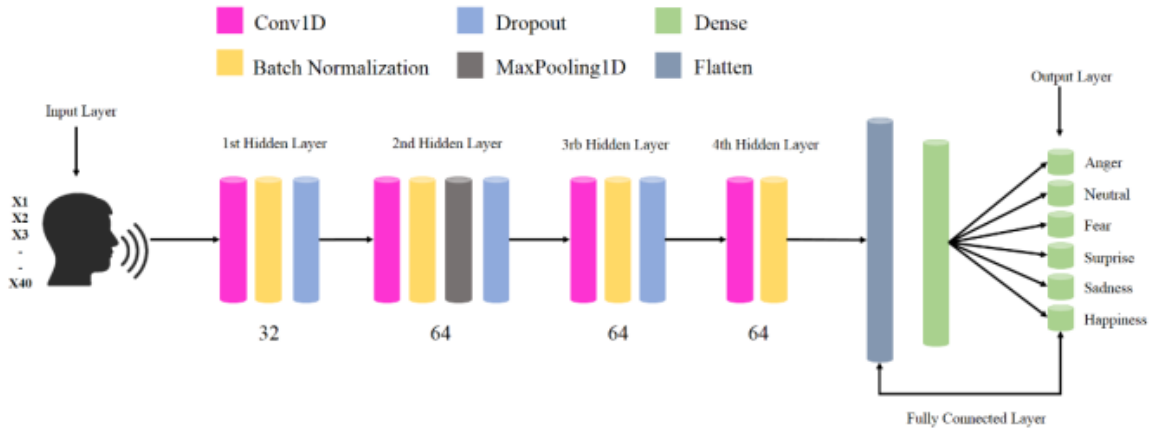

Figure 3. The proposed CNN architecture

The parameters of the CNN model that have been selected are as presented in Table 2, using input neurons of 40 neurons according to the results of MFCC feature extraction. There are 4 hidden layers, with 32, 64, 64, and 64 neurons in each layer. Relu uses input layer and hidden layer activation functions. In each hidden layer and output layer, L2 regularization of 0.0001 is implemented to reduce over-fitting and dropout of 0.07 and 0.14. The optimization used by Adam had a learning rate of 0.0001, a number of epochs of 300, and a batch size of 16 [10]. Then, ModelCheckpoint callbacks are used to store the best model based on the smallest validation loss value during the training process.

Table 2. Parameters of CNN Architecture

Parameter	Parameter Value
Input Neuron	40
Hidden Layer	4
Hidden Neuron	32, 64, 64, 64
Epoch	300
Batch Size	16
Activation Function	Relu
Dropout	0,07 dan 0,14
Optimazer	Adam
Learning rate	0,0001

2.2. Evaluation

Model evaluation will use performance metrics to determine the performance of the best classification model. The performance metrics that will be used to evaluate the model are accuracy, f1-score, precision, recall, and kappa. Accuracy is the proportion of accurately anticipated tests to the whole number of tests [21]. The formula for finding accuracy values in multi-class classification is as follows:

$$Accuracy = \sum_{k=1}^{K} \frac{(TP_k + TN_k)}{TP_k + FP_k + TN_k + FN_k} \quad (1)$$

The ratio of accurately predicted positive samples to all samples projected as positive is known as precision [21]. The formula for finding precision values in multi-class classification is as follows:

$$Macro\ Average\ Precision = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FP_k}}{\mathrm{K}} \quad (2)$$

Recall is the division of true positive (TP) components partitioned by the overall number of units classified emphatically [22]. The formula for finding recall values in multi-class classification is as follows:

$$Macro\ Average\ Recall = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}}{\mathrm{K}} \quad (3)$$

F1-Score could be a comprehensive appraise of demonstrate exactness and can be calculated as the average of precision and recall [21]. The formula for finding f1-score in multi-class classification is as follows:

$$Macro\ F1 - Score = 2\left(\frac{Macro\ Average\ Precision\ x\ Macro\ Average\ Recall}{Macro\ Average\ Precision + Macro\ Average\ Recall}\right) \quad (4)$$

Cohen's Kappa Statistics is a very good value measure that can handle the problem of multiple classes and unbalanced classes very well. The following is the formula for Cohen's Kappa on the multi-class confusion matrix [22].

$$Cohen's\ Kappa = \frac{\mathrm{c\ x\ s} - \sum_{\mathrm{k}}^{\mathrm{K}} \mathrm{p_k\ x\ t_k}}{\mathrm{s^2} - \sum_{\mathrm{k}}^{\mathrm{K}} \mathrm{p_k\ x\ t_k}} \quad (5)$$

3. RESULTS AND DISCUSSION

The experiments in this research were carried out using the Google Colabs application with the Python programming language. The Python libraries used include Keras, Scikit Learn, and Tensorflow. With the Windows 10 Pro operating system and Intel (R) Core (TM) i5 CPU M450 @ 2.40GHz and 4 Gigabite RAM.

3.1. Dataset exploration

After data exploration, it can be concluded that the dataset has six emotional classes, namely anger, fear, happiness, neutrality, sadness, and surprise. The anger class has a total of 280 audio files. The fear class has a total of 280 audio files. The happiness class has a total of 279 audio data points because there is a missing value in 1 audio data point in the recording process. In total, the neutral category consists of 280 data points. The sad category also encompasses 280 data points. At the same time, the surprise category has 280 data points. The total amount of data used is 1679, where the data used is audio data in wav format. The frequency distribution of the data used is depicted in Figure 4.

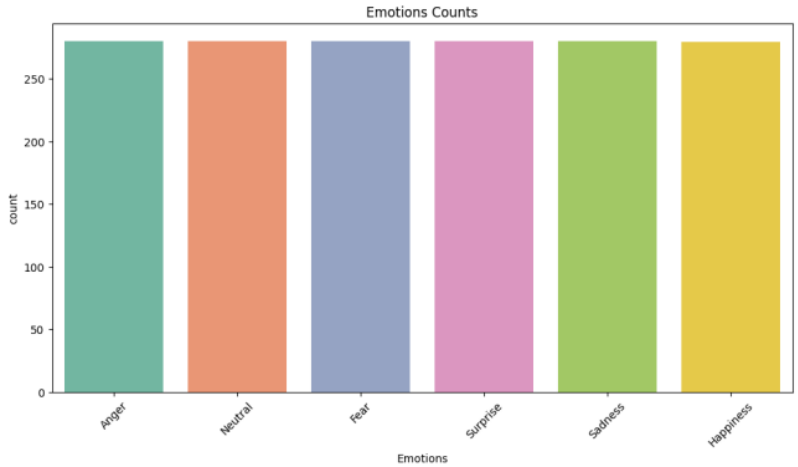

Figure 4. Frequency Distribution of Java-SED

3.2. Result of Data Preparation

The Java-SED dataset consisting of 1679 audio data is divided into 80% training with results totaling 1343 audio data and 20% testing totaling 336 audio data. Details of the data distribution for each emotion class are presented in Table 3.

Table 3. Details of Data Distribution

Emotion Class	Training Data	Testing Data
Anger	231	49
Fear	224	56
Happiness	221	58
Neutral	220	60
Sadness	225	55
Surprise	222	58

Data augmentation is applied to training data by combining the three techniques used in this research. The data augmentation techniques used are Add Noise with a noise ratio value of 0.005 [23], Time Stretch with a stretch factor value of 0.8 [24], and Pitch Shift with an n_steps value of 1 [19]. The visualization after applying the augmentation technique can be seen in Figure 5.

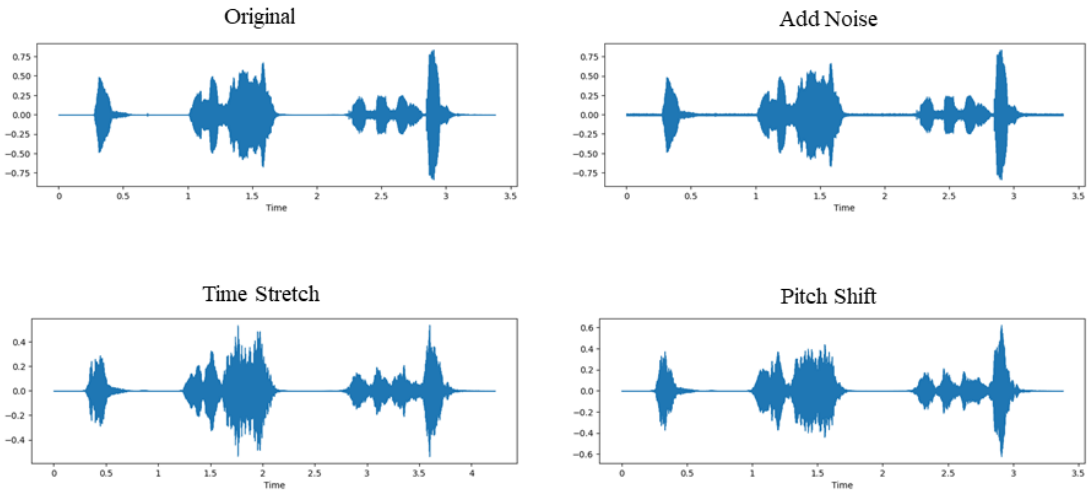

Figure 5. Before and After Implementing Data Augmentation Techniques

The application of the three data augmentation techniques above was carried out on training data of 1343 audio files. By combining these 3 data augmentation techniques, it produces 7 experimental models and 1 model using original data from the training data. The results of applying a combination of data augmentation techniques on the training data in this study are presented in Table 4.

Table 4. Results of Implementing Data Augmentation in Java-SED

Model	Data Augmentation	Addition of Data
A	Training Data	-
B	Training Data + Add Noise	1343
C	Training Data + Time Stretch	1343
D	Training Data + Pitch Shift	1343
E	Training Data + Add Noise + Time Stretch	2686
F	Training Data + Add Noise + Pitch Shift	2686
G	Training Data + Time Stretch + Pitch Shift	2686
H	Training Data + Add Noise + Time Stretch + Pitch Shift	4029

3.3. Experimental Results

Experiments were conducted on each model sequentially by implementing data augmentation techniques. The Mel Frequency Cepstral Coefficient (MFCC) feature extraction technique was employed to extract characteristics from raw human voice data in the form of digital audio. The following is a comparison of performance metrics using the MLP and CNN algorithms from the eight experimental models that have been carried out:

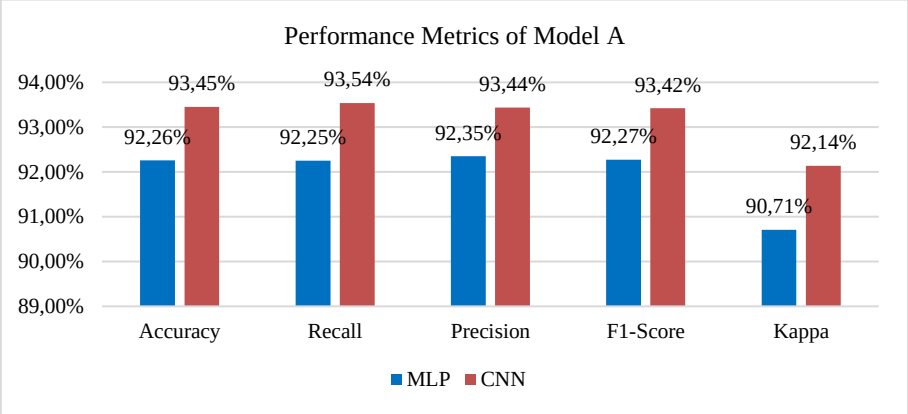

Figure 6. Comparison of Performance Metrics for Model A

Figure 6 shows that the score obtained by the CNN algorithm model in this study for performance metrics is higher than the MLP algorithm model, with an accuracy value of 93.45%, recall of 93.54%, precision of 94.44%, f1-score of 93.42%, and kappa of 92.14%.

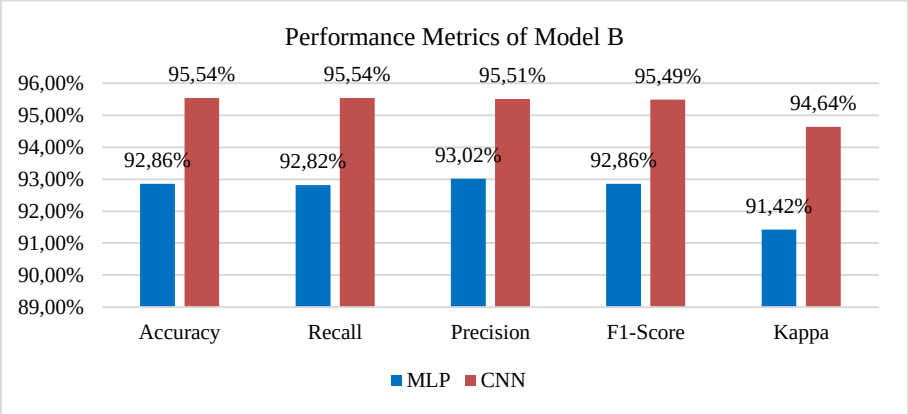

Figure 7. Comparison of Performance Metrics for Model B

Figure 7 shows that the score obtained by the CNN algorithm model in this study for performance metrics is higher than the MLP algorithm model, with an accuracy value of 95.54%, recall of 95.54%, precision of 95.51%, f1-score of 95.49%, and kappa of 94.64%.

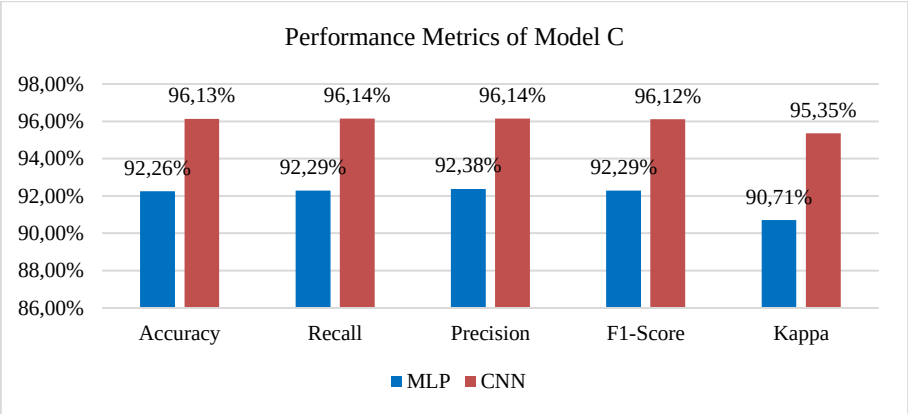

Figure 8. Comparison of Performance Metrics for Model C

Figure 8 shows that the score obtained by the CNN algorithm model in this study for performance metrics is higher than the MLP algorithm model, with an accuracy value of 96.13%, recall of 96.14%, precision of 96.14%, f1-score of 96.12%, and kappa of 95.35%.

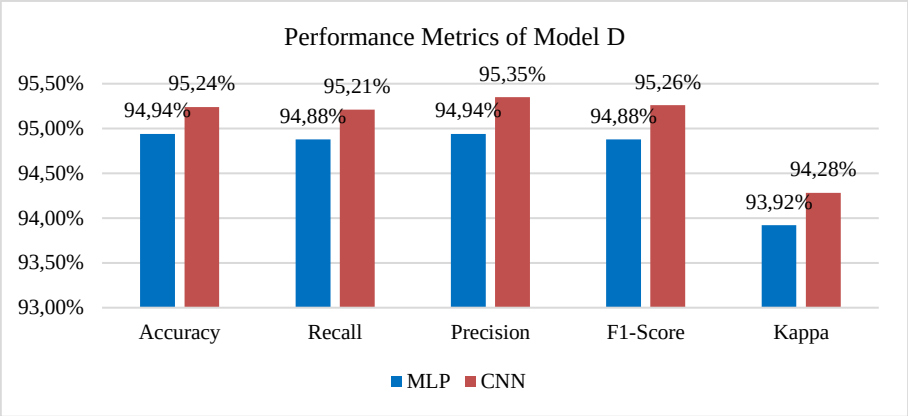

Figure 9. Comparison of Performance Metrics for Model D

Figure 9 shows that the score obtained by the CNN algorithm model in this study for performance metrics is higher than the MLP algorithm model, with an accuracy value of 95.24%, recall of 95.21%, precision of 95.35%, f1-score of 95.26%, and kappa of 94.28%.

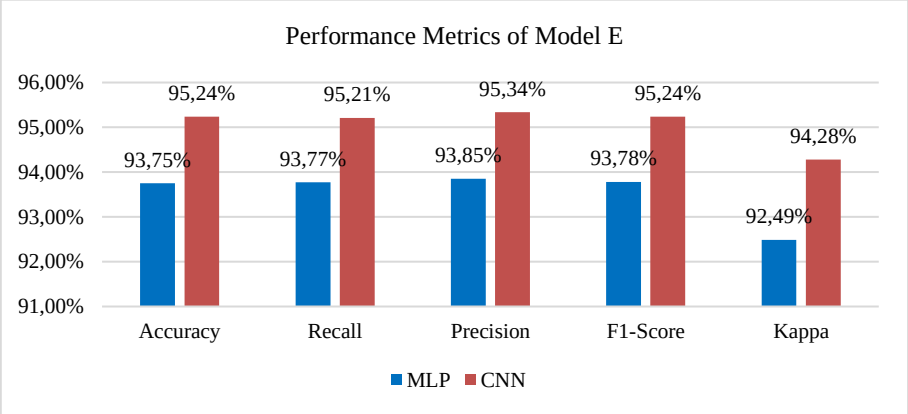

Figure 10. Comparison of Performance Metrics for Model E

Figure 10 shows that the score obtained by the CNN algorithm model in this study for performance metrics is higher than the MLP algorithm model, with an accuracy value of 95.24%, recall of 95.21%, precision of 95.34%, f1-score of 95.24%, and kappa of 94.28%.

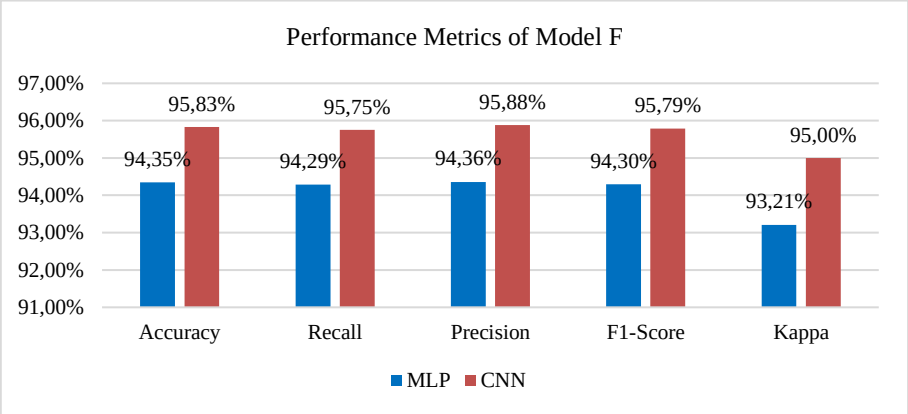

Figure 11. Comparison of Performance Metrics for Model F

Figure 11 shows that the score obtained by the CNN algorithm model in this study for performance metrics is higher than the MLP algorithm model, with an accuracy value of 95.83%, recall of 95.75%, precision of 95.88%, f1-score of 95.79%, and kappa of 95.5%.

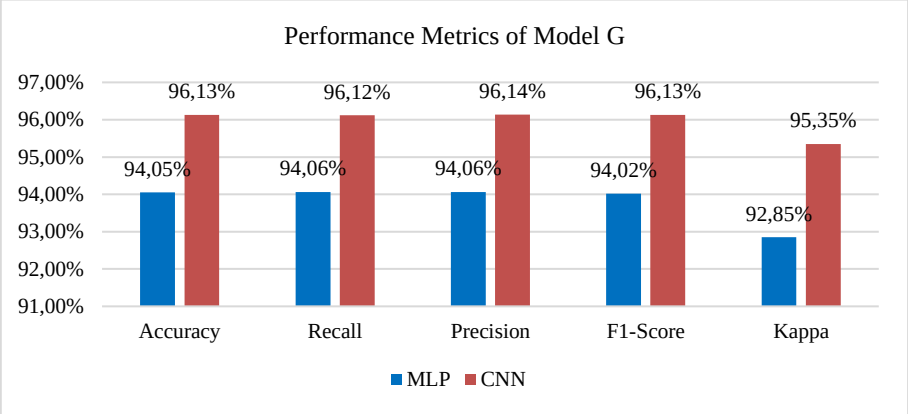

Figure 12. Comparison of Performance Metrics for Model G

Figure 12 shows that the score obtained by the CNN algorithm model in this study for performance metrics is higher than the MLP algorithm model, with an accuracy value of 96.13%, recall of 96.12%, precision of 96.14%, f1-score of 96.13%, and kappa of 95.35%.

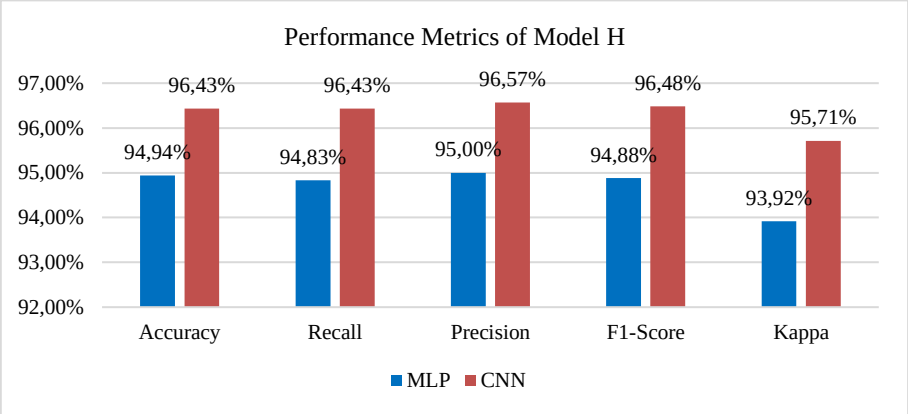

Figure 13. Comparison of Performance Metrics for Model H

Figure 13 shows that the score obtained by the CNN algorithm model in this study for performance metrics is higher than the MLP algorithm model, with an accuracy value of 96.43%, recall of 96.43%, precision of 96.57%, f1-score of 96.48%, and kappa of 95.71%.

Based on the results of the eight experimental models conducted, the comparison of performance metrics for various models (A to H) indicates that the CNN algorithm consistently outperforms the MLP algorithm in this study. Across different models, the CNN algorithm consistently achieves higher accuracy, recall, precision, f1-score, and kappa values compared to the MLP algorithm. This suggests that the CNN algorithm is more effective in producing accurate and reliable results for the given task, making it a preferable choice based on the performance metrics presented in the figures. Overall, the experimental model that yields the best performance score involves using the CNN algorithm with model H. This model incorporates data augmentation techniques such as Add Noise, Time Stretch, and Pitch Shift. The amount of data used in model H is 5,372 for training and 336 for testing. Therefore, the experiment with model H is selected as the best-proposed model for human emotion classification using the CNN algorithm, based on the performance metric results that achieved the highest score.

3.4. Discussion

The outcomes of the proposed CNN model's performance metrics were compared with previous research using the MLP model [18]. This research has similarities in the use of Mel Frequency Cepstral Coefficient (MFCC) feature extraction and the Java-SED dataset. Figure 13 shows a comparison of the accuracy score of the proposed CNN and MLP models with the MLP model from previous research.

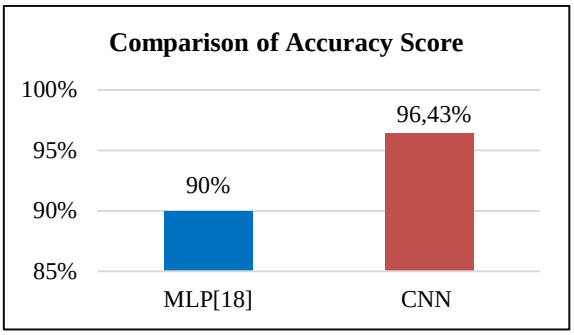

Figure 13. Comparison With Previous Research Using Java-SED

Based on the exploratory comes about in Figure 13 above, previous research on the MLP algorithm produced an accuracy of 90% with a proportion of 75% training data and 25% testing data. Then, the proposed CNN algorithm produces an accuracy of 96.43% with a proportion of 80% training data and 20% testing data. Apart from that, the proposed model also applies data augmentation techniques to the testing data, namely adding noise, time stretching, and pitch shifting to add new synthetic data. Adding new data by applying data augmentation techniques also has an impact on increasing the performance of the proposed model by 6.43%. Overall, the results obtained show that the CNN algorithm produces better performance compared to the MLP algorithm from previous research [18].

Table 5. Comparison With Previous Research Using MLP

No	Author	Dataset	Feature Extraction	Methodology	Accuracy
1	Hoseini [5]	Persian Speech	LPC Coefficients + Frequency Features	MLP	87.74%
2	Aljuhani et al. [7]	Arabic Dialect	MFCC + Mel Spectogram + Spectral Contrast	MLP	71.43%
3	Rumagit et al. [8]	Indonesia Language	mel-spectrogram, chroma, and MFCC	MLP	84.62%
4	Alnuaim et al. [9]	RAVDESS	MFCC + STFT + Mel spectrogram	MLP	81.00%
5	Arifin et al. [18]	Java-SED	MFCC	MLP	90.00%
6	This research	Java-SED	MFCC	MLP	94.94%

Based on the analysis of the table 5, it can be concluded that these studies generally employ a similar approach in sound classification methodology, namely Multilayer Perceptron (MLP). The used feature extraction methods vary, but it is evident that a majority of the studies utilize Mel-Frequency Cepstral Coefficients (MFCC), sometimes in combination with other features such as LPC Coefficients, Frequency Features, Spectrogram, and Short-Time Fourier Transform (STFT). Each study focuses on a different dataset, encompassing specific languages or dialects such as Persian Speech, Arabic dialect, Indonesia Language, RAVDESS, and Java-SED. The latest research, stemming from this study, achieved the highest accuracy rate of 94.94% for the Java-SED dataset. These findings suggest that the utilization of MFCC features with the MLP method can deliver excellent performance in sound recognition, particularly for specific languages or dialects. Although other studies also achieved good accuracy, this recent research stands out by attaining superior results, showcasing the potential of this method in the context of sound recognition on the Java-SED dataset.

Table 6. Comparison With Previous Research Using CNN

No	Author	Dataset	Feature Extraction	Methodology	Accuracy
1	Patel et al. [1]	TESS	MFCC	CNN 1D	96.00%
2	ozer [3]	EMOVO	Pseudo-colored log-power rate map	CNN 2D	68.93%
3	Issa et al. [4]	IEMOCAP	MFCC + Chromagram + Mel Spectogram + Spectral Contrast + Tonnetz	CNN 1D	64.30%
4	Jahangir et al. [10]	EMO-DB	MFCC + Mel Spectogram + Spectral Contrast + Tonnetz + Chromagram + Δ MFCC + ΔΔ MFCC	CNN 1D	96.70%
5	Jahangir et al. [10]	RADVESS	MFCC + Mel Spectogram + Spectral Contrast + Tonnetz + Chromagram + Δ MFCC + ΔΔ MFCC	CNN 1D	90.60%
6	Jahangir et al. [10]	SAVEE	MFCC + Mel Spectogram + Spectral Contrast + Tonnetz + Chromagram + Δ MFCC + ΔΔ MFCC	CNN 1D	93.20%
7	Alnuaim et al. [12]	BAVED	RMSE + Energy + ZCR + Flux + Centroid + Roll off + Chroma + Mel Spectrogram + MFCC	CNN 1D	97.09%
8	Alnuaim et al. [12]	ANAD	RMSE + Energy + ZCR + Flux + Centroid + Roll off + Chroma + Mel Spectrogram + MFCC	CNN 1D	96.44%
9	This research	Java-SED	MFCC	CNN 1D	96.43%

Table 6 presents a comprehensive summary of the research outcomes from various scholars in the field of voice or emotion recognition. From the table analysis, it can be inferred that researchers employ diverse feature extraction methods, including MFCC, Chromagram, Mel Spectrogram, and various others. The most commonly adopted methodology is the Convolutional Neural Network (CNN), in both 1D and 2D formats. CNN has an advantage in handling spatial or temporal data, making it effective in analyzing complex sound patterns, particularly in the context of emotion recognition. CNN has proven to yield excellent results in several studies, such as Alnuaim et al.'s research on the BAVED dataset achieving an accuracy of 97.09% [12], and Jahangir et al.'s study on the EMO-DB dataset reaching 96.70% accuracy [10]. The recent study included in this table also achieved satisfactory results with an accuracy of 96.43% on the Java-SED dataset. The strength of CNN is in its capability to automatically derive hierarchical features from intricate data, making it highly effective in sound pattern recognition tasks. Despite variations in datasets and feature extraction methods, some studies consistently achieve high accuracy levels, reflecting the success of the CNN approach in overcoming the challenges of sound recognition.

4. CONCLUSION

By detailing the results of research analysis and discussion, it can be concluded that the use of data augmentation techniques has the capability to enhance the performance of algorithmic models employed in human emotion classification. Based on the evaluation results, the proposed CNN algorithm model obtained the highest performance metric score of 96.43% for accuracy, recall of 96.43%, precision of 96.57%, F1 score of 96.48%, and kappa of 95.71% by applying Add Noise, Time Stretch, and Pitch Shift techniques. The proposed classification algorithm model produces better performance compared to the performance obtained by the multilayer perceptron algorithm carried out in previous research [18].

REFERENCES

[1] N. Patel, S. Patel, and S. H. Mankad, "Impact of autoencoder based compact representation on emotion detection from audio," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 2, pp. 867–885, 2021, doi: 10.1007/s12652-021-02979-3.

[2] H. Ibrahim, C. K. Loo, and F. Alnajjar, "Bidirectional parallel echo state network for speech emotion recognition," *Neural Comput. Appl.*, vol. 34, no. 20, pp. 17581–17599, 2022, doi: 10.1007/s00521-022-07410-2.

[3] I. OZER, "Pseudo-colored rate map representation for speech emotion recognition," *Biomed. Signal Process. Control*, vol. 66, no. February, p. 102502, 2021, doi: 10.1016/j.bspc.2021.102502.

[4] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, pp. 1–11, 2020, doi: 10.1016/j.bspc.2020.101894.

[5] S. M. Hoseini, "Persian Speech Emotion Recognition Approach based on Multilayer Perceptron," *Int. J. Digit. Content Manag.*, vol. 2, no. 3, pp. 177–187, 2021, doi: https://doi.org/10.22054/dcm.2021.13682.

[6] A. P. Wibawa, W. Lestari, A. B. P. Utama, I. T. Saputra, and Z. N. Izdihar, "Multilayer Perceptron untuk Prediksi Sessions pada Sebuah Website Journal Elektronik," *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 57–67, 2020, doi: 10.33096/ijodas.v1i3.15.

[7] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic Speech Emotion Recognition from Saudi Dialect Corpus," *IEEE Access*, vol. 9, pp. 127081–127085, 2021, doi: 10.1109/ACCESS.2021.3110992.

[8] R. Y. Rumagit, G. Alexander, and I. F. Saputra, "Model Comparison in Speech Emotion Recognition for Indonesian Language," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 789–797. doi: 10.1016/j.procs.2021.01.098.

[9] A. A. Alnuaim *et al.*, "Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier," *J. Healthc. Eng.*, vol. 2022, pp. 1–12, 2022, doi: 10.1155/2022/6005446.

[10] R. Jahangir, Y. W. Teh, G. Mujtaba, R. Alroobaea, Z. H. Shaikh, and I. Ali, "Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion," *Mach. Vis. Appl.*, vol. 33, no. 41, pp. 1–16, 2022, doi: 10.1007/s00138-022-01294-x.

[11] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, "Heart sound classification based on improved MFCC features and convolutional recurrent neural networks," *Neural Networks*, vol. 130, pp. 22–32, 2020, doi: 10.1016/j.neunet.2020.06.015.

[12] A. A. Alnuaim *et al.*, "Human-Computer Interaction with Detection of Speaker Emotions Using Convolution Neural Networks," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–16, 2022, doi: 10.1155/2022/7463091.

[13] S. Sultana, M. Z. Iqbal, M. R. Selim, M. Rashid, and M. S. Rahman, "Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks," *IEEE Access*, vol. 10, pp. 564–578, 2022, doi: 10.1109/ACCESS.2021.3136251.

[14] O. S. Ghongade, S. K. S. Reddy, Y. C. Gavini, S. Tokala, and M. K. Enduri, "Acute Lymphoblastic Leukemia Blood Cells Prediction Using Deep Learning & Transfer Learning Technique," *Indones. J. Electr. Eng. Informatics*, vol. 11, no. 3, pp. 778–790, 2023, doi: 10.52549/ijeei.v11i3.4855.

[15] R. Nurcahyo and M. Iqbal, "Pengenalan Emosi Pembicara Menggunakan Convolutional Neural Networks," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 1, pp. 115–122, 2022, doi: 10.29207/resti.v6i1.3726.

[16] M. R. Ahmed, S. Islam, A. K. M. M. Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," *Expert Syst. Appl.*, vol. 218, pp. 1–21, 2023, doi:

10.1016/j.eswa.2023.119633.
[17] N. T. Pham *et al.*, "Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition," *Expert Syst. Appl.*, vol. 230, pp. 1–13, 2023, doi: 10.1016/j.eswa.2023.120608.
[18] F. Arifin, A. S. Priambodo, A. Nasuha, A. Winursito, and T. S. Gunawan, "Development of Javanese Speech Emotion Database (Java-SED)," *Indones. J. Electr. Eng. Informatics*, vol. 10, no. 3, pp. 584–591, 2022, doi: 10.52549/ijeei.v10i3.3888.
[19] S. Wei, S. Zou, F. Liao, and W. Lang, "A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, 2020, pp. 1–8. doi: 10.1088/1742-6596/1453/1/012085.
[20] A. A. C. Alves *et al.*, "Integrating Audio Signal Processing and Deep Learning Algorithms for Gait Pattern Classification in Brazilian Gaited Horses," *Front. Anim. Sci.*, vol. 2, no. August, pp. 1–19, 2021, doi: 10.3389/fanim.2021.681557.
[21] S. K. Challa, A. Kumar, and V. B. Semwal, "A multibranch CNN-BiLSTM model for human activity recognition using wearable sensor data," *Vis. Comput.*, vol. 38, pp. 4095–4109, 2021, doi: 10.1007/s00371-021-02283-3.
[22] M. Grandini, E. Bagli, and G. Visani, "METRICS FOR MULTI-CLASS CLASSIFICATION: AN OVERVIEW," *Arxiv*, vol. abs/2008.0, pp. 1–17, 2020, doi: 10.48550/arXiv.2008.05756.
[23] A. A. Abdelhamid *et al.*, "Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm," *IEEE Access*, vol. 10, pp. 49265–49284, 2022, doi: 10.1109/ACCESS.2022.3172954.
[24] A. C. Shruti, R. H. Rifat, M. Kamal, and M. G. R. Alam, "A Comparative Study on Bengali Speech Sentiment Analysis Based on Audio Data," in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, IEEE, 2023, pp. 219–226. doi: 10.1109/BigComp57234.2023.00043.

BIOGRAPHY OF AUTHORS

Muji Ernawati was born in Kebumen in 1998. She has been pursuing a master's degree in computer science at Universitas Nusa Mandiri, Indonesia, with her specialization being data mining. Currently, she is joining the Universitas Nusa Mandiri Computer Science Community (UCSC) as treasurer. She has published several research papers in national journals and conferences related to machine learning and deep learning. She is interested in research related to machine learning, deep learning, and natural language processing.

Prof. Dr. Ir. Dwiza Riana, S.Si, MM, M.Kom, IPU, ASEAN.Eng has been a permanent lecturer at the Faculty of Information Technology, Computer Science Study Program at Universitas Nusa Mandiri since 2003. Currently she serves as Chancellor of Universitas Nusa Mandiri. Completed Doctoral Education (S3) in the Electrical and Informatics Engineering Study Program at the Bandung Institute of Technology (ITB) in 2015. Active on the DKI Jakarta Province Aptikom Advisory Board, as Provincial Aptikom Advisory Board West Java, Central Aptikom Management, APTIKOM Journal Publishing Team, as administrator of LAM Infocom Division I, Management of the Association of Indonesian Private Universities (APTISI) Region III DKI Jakarta for the 2022-2026 period and as Vice Chair of IEEE, Computational Intelligence Society, Indonesia Chapter. She skilled in Computer Science, Image Processing, Data Mining, Public Speaking, Management, Information Systems, and Machine Learning.