

# Classification of regional language dialects using convolutional neural network and multilayer perceptron

Fahmi B. Marasabessy<sup>1</sup>, Dwiza Riana<sup>2</sup>, Muji Ernawati<sup>3</sup>

<sup>1</sup>Computer Science Study Program, Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, Indonesia

<sup>2</sup>Doctoral Program in Informatics, Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, Indonesia

<sup>3</sup>Informatics Study Program, Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, Indonesia

## Article Info

### Article history:

Received Dec 4, 2024

Revised Sep 11, 2025

Accepted Oct 16, 2025

### Keywords:

Convolutional neural network

Dialect recognition

Mel-frequency cepstral coefficients

Multilayer perceptron

Regional language dialects

## ABSTRACT

Regional languages are vital for communication and preserving cultural identity, safeguarding local heritage. However, globalization and modernization endanger their existence as they are increasingly replaced by national or global languages. Despite progress in dialect recognition research, particularly for certain languages, further studies are needed to improve model performance and address less-represented dialects, including those in Indonesia. This study enhances a custom-built dataset for dialect recognition through the application of data augmentation techniques, specifically adding noise, time stretching, and pitch shifting. Using Mel-frequency cepstral coefficients (MFCC) for feature extraction, it evaluates the performance of convolutional neural network (CNN) and multilayer perceptron (MLP) in classifying six Indonesian dialects. Results indicate that CNN outperformed, achieving 97.92% accuracy, 97.90% recall, 97.97% precision, 97.92% F1-score, and a kappa score of 97.49% with combined augmentation techniques, setting a foundation for further research.

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



## Corresponding Author:

Dwiza Riana

Doctoral Program in Informatics, Faculty of Information Technology, Universitas Nusa Mandiri

Jatiwaringin Road No. 2, Cipinang Melayu, Makasar, East Jakarta, DKI Jakarta, Indonesia

Email: dwiza@nusamandiri.ac.id

## 1. INTRODUCTION

Globalization has had a significant impact on the preservation of regional languages and local cultural identities, particularly in Indonesia, a country known for its rich linguistic and cultural diversity. According to a report by UNESCO (2021), around 40 percent of the world's languages are endangered, mainly due to the lack of intergenerational transmission. A similar trend is evident in Indonesia, where the use of regional languages among younger generations is increasingly being replaced by Indonesian or foreign languages, especially English, in daily activities such as in education, the workplace, and social media. This shift raises concerns about the potential extinction of regional languages and the cultural values embedded within them [1].

Indonesia ranks as the fourth most populous country in the world, with approximately 273 million people spread across 17,508 islands. More than 700 languages are spoken throughout the archipelago, highlighting its vast linguistic diversity. Unfortunately, many of these languages are now at risk. According to data from Ethnologue, 440 local languages in Indonesia are classified as endangered, and 12 have become extinct. A study of 98 local languages found that nearly half are considered endangered, while another study reported that 71 out of 151 local languages have fewer than 100,000 speakers [2]. Language is closely associated with dialects and accents, all of which are influenced by factors such as the environment, language proficiency, and social interaction [3]. A dialect is a variation of a language spoken by a specific group of speakers, characterized by unique traits that differ from one region to another [4]. According to the study on

the dynamics of language interaction in multicultural urban communities, each language possesses inherent uniqueness that reflects the cultural identity of its speakers. Additionally, every region exhibits distinctive linguistic characteristics and traditions, which evolve through continuous social interaction and cultural blending among diverse groups [5]. From a linguistic perspective, approaches to the concept of dialect can vary among linguists. Dialects encompass differences not only in phonology, lexicon, or grammar but also in pronunciation and everyday language use [6]. Therefore, preserving regional languages is essential to maintaining cultural identity and safeguarding local heritage.

For foreign speakers learning Indonesian, communication success is often measured by their ability to converse with native speakers. However, challenges arise when they interact directly with native speakers due to variations in local dialects. According to Wang [7], dialects are also difficult to be accurately understood by speech recognition systems because of their unique pronunciation, vocabulary, and grammatical structure. To address these issues, dialect recognition has increasingly been integrated as an essential component within voice recognition technology, enabling systems to process regional linguistic diversity more effectively [8]. Voice recognition technology offers a potential tool for supporting the preservation and understanding of regional languages, particularly through dialect classification. Research on voice recognition technology related to dialect or accent identification in specific countries or regions has been conducted extensively. Examples include recognizing Kurdish dialects using 1D convolutional neural networks (CNN) [9] and identifying differences between two Colombian dialects, "Antioqueño" and "Bogotano," using CNN [10]. Similarly, in Indonesia, there have been studies focused on dialect or regional language classification. For example, Tawaqal and Suyanto [11] used a deep recurrent neural network (DRNN) to identify five main dialects: Javanese, Sundanese, Banjar, Buginese, and Malay. In addition, Nugroho *et al.* [12] developed a data augmentation approach combined with a seven-layer deep neural network (DA-DNN7L) to classify ethnic speakers using 700 utterances from 70 ethnic groups. Other studies have also addressed dialect recognition in Indonesian languages, such as the detection of Sundanese [13] and Balinese Badung [14]. Dialect identification involves determining the dialect category of spoken utterances. This task focuses on recognizing the speaker's regional dialect within a particular language based solely on the available acoustic signals [15].

This study aims to identify differences among various regional dialects in Indonesia through voice analysis. A custom dataset was developed, consisting of six classes representing dialects from Medan, Minang, Sunda, Lombok, Madura, and Ambon. While the dataset provides a foundation for exploring these dialects, it represents only a subset of Indonesia's rich linguistic diversity, warranting further expansion in future research. The research builds upon prior work [16] by implementing previously proposed techniques, including data augmentation methods (such as adding noise, time stretching, and pitch shifting), Mel-frequency cepstral coefficients (MFCC) for feature extraction, and comparing the performance of CNN and multilayer perceptron (MLP) models. These techniques were applied to the newly constructed six-class dataset to evaluate the consistency of their performance on new data. By leveraging these methods, the study aimed to determine whether the algorithms could maintain high accuracy when applied to a different dataset, effectively enabling the classification and identification of dialects based on voice features.

This paper is structured as follows: section 2 discusses the stages and methods applied in this study. The conditions and results of the experiments are presented in section 3. Finally, the conclusion of this research is provided in section 4.

## 2. METHOD

In the research methodology chapter, the process or scientific method used to obtain data for research purposes is defined. This method includes scientific approaches, steps, and types, as well as the limitations of the scientific method. Figure 1 illustrates the stages of the research that will be conducted in this study.

### 2.1. Dataset

A dataset is a collection of data that provides an overview of a specific topic [17]. The dataset used in this research is a private dataset, named Indonesian dialects dataset. It consists of dialects from several regional languages in Indonesia, namely Medan, Minang, Sunda, Lombok, Madura, and Ambon. The dataset contains a total of 1,996 files in AAC format [17]. This dataset was collected using a smartphone and a wireless microphone over a period of 2 months, and the files or words and sentences used are entirely the author's own, which the author compiled within 1 week. The dataset features six speakers aged between 30 and 50 years old, with two female and four male speakers [18]. The speakers are individuals who still fluently use their regional languages, complete with the local accents. Each speaker will deliver an average of 200 words or sentences that have been pre-prepared by the author. Thus, the database contains a total of 1,996 utterances, with each dialect averaging 200 words/sentences. The number of words consists of 100 nouns or adjectives, 50 interrogative sentences, and 50 imperative sentences, as selected for this dataset, as shown in Table 1.

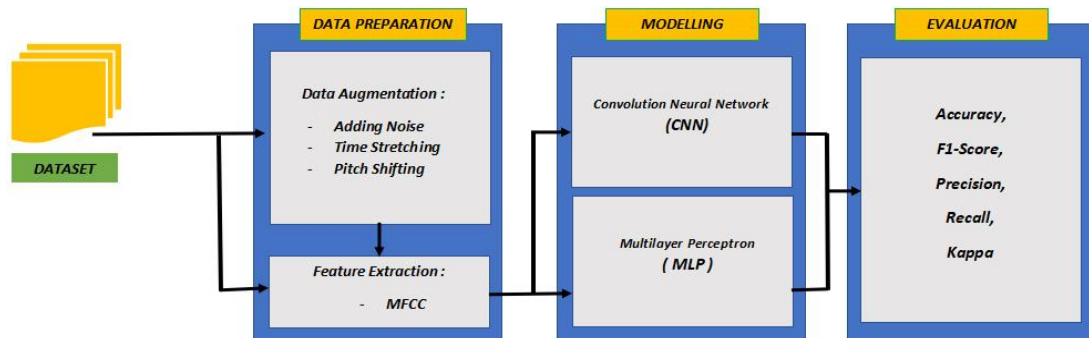


Figure 1. Stages of research

Table 1. Vocabulary for the dataset in dialect languages

Words	Imperative sentences	Interrogative sentences
Father	Please don't be noisy, kids!	Do you like spicy food?
Sibling	No smoking allowed!	How do you feel about your current job?
Lion	Please make black coffee for father!	When are you planning to leave?
River	Do not speak without permission in the meeting room!	Who is making noise from the next room?
Hat	No using mobile devices while driving!	Why do you always get angry every time I meet you?

## 2.2. Data preparation

Data preparation involves converting raw data into a more practical and efficient form, ensuring that the dataset to be processed and analyzed is both accurate and consistent. In this phase, the dialect dataset is first divided, and then data augmentation and feature extraction are applied to the training data to make it suitable for the proposed model. Prior to feeding the dataset into the model algorithm, data normalization is carried out using standard scaler, and dialect classes are encoded with one hot encoding to optimize the model's data processing. In the initial pre-processing stage, data splitting is performed to divide the dataset into two parts: training data and testing data. 80% of the training data will be used to train the developed model, while the remaining 20% will be used as validation and testing data to evaluate the model's performance based on the established performance metrics. The random sampling method will be applied to ensure that the data modeling process is not biased by any potential feature differences in the dataset.

In this pre-processing stage, data augmentation will be applied to the training data by adding several data augmentation techniques to generate new synthetic data. The data augmentation methods to be applied include: adding noise with a noise ratio of 0.005, applying time stretching with a factor of 0.8, and shifting the pitch with *n\_steps* set to 1 [16]. After the raw data is processed through augmentation, the next step is to extract data by applying the feature extraction method on the MFCC from the augmented data to detect language dialects. MFCC is a method used to decompose speech signals into components to represent information about pitch and vocal tract characteristics [19]. This technique simulates human auditory behavior by distinguishing sound frequencies, with frequency bands calculated logarithmically [20]. The processes used in feature extraction at the MFCC layer include pre-emphasize, frame blocking, windowing process, fast fourier transform (FFT), Mel frequency warping, discrete cosine transform (DCT), and cepstral liftering.

## 2.3. Modeling

The experiments involve combining three data augmentation techniques to examine the effect of incorporating additional data into the training set on the classification model's performance. Table 2 outlines the experimental models to be conducted in this study, which are based on the combination of adding noise, time stretching, and pitch shifting techniques. The original data in Table 2 refers to the testing data that has been split during the data preparation stage.

Experiments will be conducted on eight models, each of which implements the MFCC feature extraction technique. These models are designed to evaluate the impact of different data augmentation combinations on classification performance. This study will conduct dialect classification modeling using the CNN and MLP algorithms adopted from previous studies [16].

The architecture of the MLP model will be compared with the proposed CNN model architecture. MLP is known for its strong scalability and efficiency in learning patterns compared to other classifiers, thanks to its compact structure and adaptive mechanisms [21]. The MLP model used has two hidden layers, with 488 nodes in the first layer and 443 nodes in the second. The MLP architecture will be applied to the eight experimental models with a random state parameter set to 1, a maximum number of iterations of 325, a learning

rate of 0.00233,  $n\_iter\_no\_change = 27$ , and an alpha value of 0.00185 for L2 regularization. The variation of the MLP architecture with parameters and their values is an architecture that has been carried out by previous studies [22].

The architecture of this model was developed using a CNN implemented with the Keras library in Python. CNN is a variation of the MLP inspired by the human neural network [23]. In this study, the proposed architecture employs an 1D CNN. The configuration of the model consists of 40 input neurons, four hidden layers with 32, 64, 64, and 64 hidden neurons respectively, and is trained for 300 epochs with a batch size of 16. The rectified linear unit (ReLU) activation function is applied, along with dropout rates of 0.07 and 0.14 to prevent overfitting. The model optimization uses the Adam optimizer with a learning rate of 0.0001 [16].

The next step is to build the CNN architecture based on the model parameters that have been selected and proposed for the classification model. The construction of the CNN architecture is presented in Figure 2. This figure provides a visual representation of the layers and configurations applied in the model.

Table 2. Experimental setup [16]

Model	Data augmentation	Feature extraction
A	Training data	MFCC
B	Training data + adding noise	MFCC
C	Training data + time stretching	MFCC
D	Training data + pitch shifting	MFCC
E	Training data + adding noise + time stretching	MFCC
F	Training data + adding noise + pitch shifting	MFCC
G	Training data + time stretching + pitch shifting	MFCC
H	Training data + adding noise + time stretching + pitch shifting	MFCC

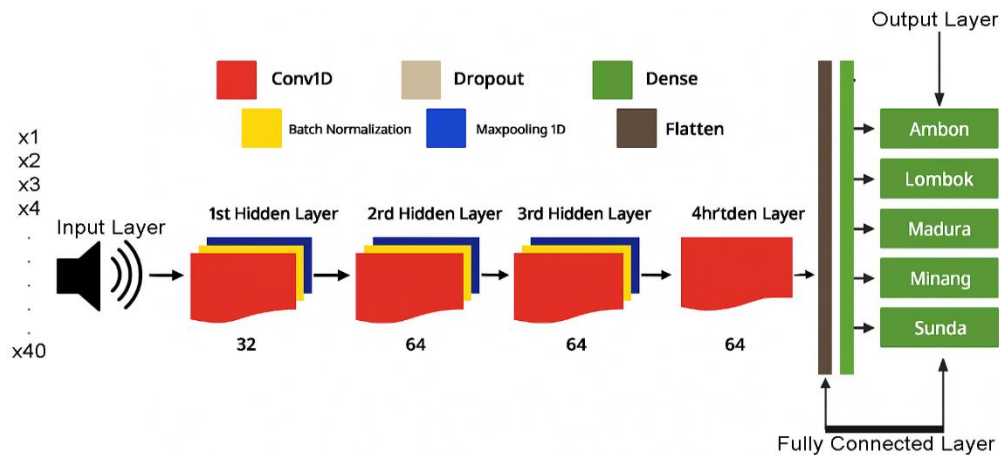


Figure 2. Convolutional neural network architecture

This study adopts an architecture based on previous research [16], as illustrated in Figure 2. The design features a one-dimensional CNN for classifying human dialects using features extracted from audio files. This 1D CNN design accepts input data in the form of a  $40 \times 1$  array, corresponding to features extracted from the dialect dataset. The model consists of four hidden layers, each with a kernel size of 3, configured with 32, 64, 64, and 64 neurons, respectively. The activation function for the input and hidden layers is ReLU. To address overfitting, the model incorporates batch normalization, L2 regularization (value: 0.0001), and dropout rates of 0.07 and 0.14 applied to specific layers. The architecture also includes a max pooling layer with a pool size of 2 and a fully connected output layer with six units, corresponding to the six dialect classes. The output layer uses the SoftMax activation function for classification. Optimization is carried out using the Adam algorithm with a learning rate of 0.0001, and categorical cross-entropy is used as the loss function. The model is trained for 300 epochs with a batch size of 16, and a ModelCheckpoint callback is used to save the best-performing model based on the lowest validation loss during training.

#### 2.4. Evaluation

After conducting the experiments, the next step is to evaluate each experimental model using CNN and MLP. The evaluation will be performed using testing data. The evaluation process will utilize the

Scikit-learn libraries. The performance metrics used for model evaluation are accuracy, F1-score, precision, recall, and kappa. The selection of the proposed algorithm model will be determined based on the performance metrics, with the best results from the eight experimental models in Table 2. Furthermore, in the CNN algorithm, to ensure that the selected model is not overfitting, the experiment process includes plotting training and testing based on the history of accuracy and loss. The formula for finding the measurement metric value used in multiclass classification is as follows [16]:

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

In (1), it measures the proportion of overall correct predictions to the total predictions.  $K$  is the number of classes;  $TP_k$  is the true positive for class  $k$ ;  $TN_k$  is the true negative;  $FP_k$  is the false positive; and  $FN_k$  is the false negative.

To evaluate model performance more comprehensively across all classes, several additional metrics are used. In (2) defines the macro average precision, which calculates the average precision across all classes by taking the mean of individual class precision scores. In (3) represents the macro average recall, which averages the recall values for each class, reflecting the model's ability to correctly identify positive instances.

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

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

The balance between these two metrics is captured through the macro F1-score as shown in (4), which provides a harmonic mean between precision and recall to ensure fairness in multi-class evaluation. Furthermore, in (5) presents Cohen's Kappa ( $\kappa$ ) measures the agreement between the model predictions and the actual labels while considering the possibility of agreement occurring by chance.  $c$  is the number of correct predictions (diagonal elements of the confusion matrix),  $s$  is the total number of samples,  $p_k$  is the number of predictions for class  $k$ , and  $t_k$  is the actual number of samples for class  $k$ .

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

$$\text{Cohen's Kappa} = \frac{c \times s - \sum_{k=1}^K p_k \times t_k}{s^2 - \sum_{k=1}^K p_k \times t_k} \quad (5)$$

### 3. RESULTS AND DISCUSSION

The system components used for the experiments in this research are specified as follows: the operating system is Windows 10 Home Single Language, running on an Intel(R) Core(TM) i3-3120M CPU @2.50 GHz processor. The system is equipped with 6 GB of RAM and a 298 GB hard disk. For data mining applications, Google Colab was utilized to support the experimentation and analysis process.

#### 3.1. Dataset exploration

The first step conducted in the experiment of this research was performing a statistical analysis of the dataset through the exploratory data analysis (EDA) process. The dataset consists of six dialect language classes: Ambon, Lombok, Madura, Medan, Minang, and Sunda. After exploring the data, the Ambon class contains 200 audio data, the Lombok class has 200 audio data, and the Madura class contains 197 audio data due to the voice recording process. The Medan class has 200 audio data, the Minang class contains 199 audio data, and the Sunda class has 200 audio data. The total number of data used is 1,996, all of which are audio data in AAC format. However, the total number of datasets, which should have been 1,200, decreased because some data were lost during the transfer process from the smartphone to the laptop/computer.

#### 3.2. Result of data preparation

Before conducting experiments for modeling, the dataset undergoes a data preparation stage. This phase is essential to ensure that the data is ready for processing by the classification models. It includes procedures such as splitting the dataset and applying augmentation techniques.

The dataset has been divided into training and testing data, where the training data is used to train the model, and the testing data is used to evaluate its performance. Out of the total 1,996 audio data in the language

dialect dataset, 80% or 956 audio data were allocated for training, while 20% or 240 audio data were used for testing. The distribution of data across the six dialect classes is relatively balanced. For the training set, Ambon contributed 159 samples, Lombok 165 samples, Madura 162 samples, Medan 150 samples, Minang 161 samples, and Sunda 159 samples. Meanwhile, for the testing set, Ambon consisted of 41 samples, Lombok 35 samples, Madura 35 samples, Medan 50 samples, Minang 38 samples, and Sunda 41 samples. This distribution ensures that each dialect is fairly represented in both the training and testing phases.

Data augmentation is a technique used to generate new data from existing samples. In this study, three augmentation methods were applied to enhance the training set: adding noise, time stretching, and pitch shifting. These methods help increase the diversity of training data and improve model generalization.

Figure 3 presents a comparative visualization of waveform representations before and after the application of data augmentation techniques to the audio signals. In particular, Figure 3(a) depicts the original waveform corresponding to an Ambon dialect sample, serving as the baseline reference. Figure 3(b) illustrates the waveform following the introduction of additive Gaussian noise with an amplitude parameter of 0.005, which visibly increases signal variability. Figure 3(c) demonstrates the impact of time stretching, evidenced by the elongation of the waveform duration to approximately 2.8 seconds while preserving temporal patterns. Figure 3(d) displays the waveform after applying pitch shifting, which alters the frequency content without changing the temporal scale. The red annotations in figure highlight specific regions where notable differences in amplitude and frequency characteristics are observable between the original and augmented signals.

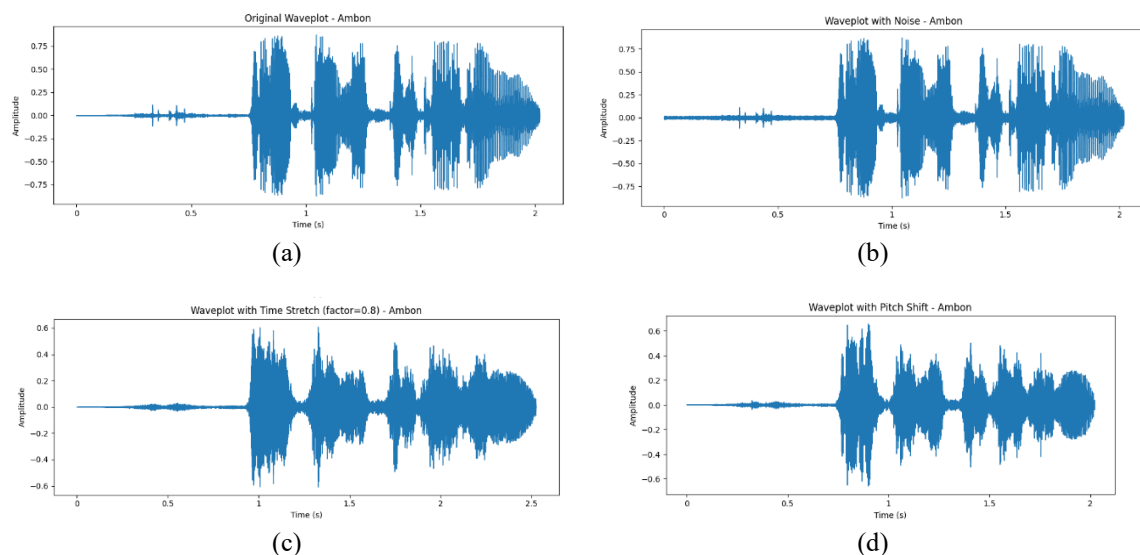


Figure 3. Waveplot visualization before and after applying data augmentation techniques: (a) original waveplot sample from the Ambon dialect, (b) waveplot after adding noise, (c) waveplot after applying time stretching, and (d) waveplot after applying pitch shifting

Applying the three data augmentation techniques to 956 audio samples for model training produced a total of seven experimental models, in addition to one model that utilized only the original training data. Experimental model A uses the 956 original audio data without augmentation. Model B applies the add noise technique, increasing the dataset to 1,912 samples. Model C applies the time stretch technique, also producing 1,912 samples, while model D applies the pitch shift technique, resulting in the same number of samples. Model E combines add noise and time stretch techniques, expanding the dataset to 2,868 samples. Similarly, model F applies add noise with pitch shift, and model G applies time stretch with pitch shift, both yielding 2,868 samples. Finally, model H integrates all three techniques (add noise, time stretch, and pitch shift), producing the largest dataset with 3,824 audio samples.

Feature extraction was performed after the data augmentation process. To generate MFCC, the Librosa library in Python was used. MFCC is designed to reflect human perception of frequency by converting conventional frequency to the Mel scale. The process begins with a pre-emphasis stage to preserve high frequencies that are usually lost during sound production. The audio signal is then divided into 40 frames. Each frame is weighted using a window function, with the audio signal split into overlapping windows based on the parameters  $n\_fft$  of 2048 and  $hop\_length$  of 512. Next, the FFT is applied to convert each frame from the time

domain to the frequency domain. The FFT values obtained are used to calculate spectral energy density, which is then mapped to a filter bank. The energy of the filter bank is calculated by multiplying the energy spectrum with the filter bank, then summing the coefficients. After the filter bank is computed, it is scaled to the Mel scale to obtain the Mel-scaled filter bank. Finally, a DCT is applied to the scaled filter bank, retaining a number of coefficients and discarding the rest. The final result is a feature vector with 40 MFCC coefficients representing the audio information.

### 3.3. Experimental results

Without applying data augmentation techniques, both CNN and MLP demonstrated limited performance, with accuracy and other metrics remaining suboptimal. The application of augmentation techniques, such as adding noise, time stretching, and pitch shifting, resulted in notable improvements in model performance, particularly for the CNN algorithm. Model H, which combines all three techniques, achieved the highest evaluation metrics on the current dataset. However, while these findings suggest that combining augmentation techniques can enhance performance, their effectiveness may depend on the dataset characteristics, and further validation on larger, more diverse datasets is necessary to confirm their general applicability.

For the MLP algorithm, the highest performance was achieved by model F, which combined add noise and pitch shift augmentation techniques. Based on Table 3, the performance comparison between CNN and MLP indicates that the application of data augmentation techniques significantly improved both models. However, the differing responses of each algorithm to augmentation techniques highlight the importance of selecting and applying the appropriate augmentation techniques tailored to the model's requirements.

Model H demonstrated the highest evaluation metrics in this study for regional dialect classification using the CNN algorithm. By applying dialect feature extraction and combining all three augmentation techniques, the model utilized 3,824 training samples and 240 testing samples. The evaluation metrics, including an accuracy of 97.92%, recall of 97.9%, precision of 97.97%, F1-score of 97.92%, and Cohen's Kappa score of 97.49%, suggest that the model performs well under the controlled conditions of this study. However, given the dataset's limited size and scope, further research is needed to validate the model's adaptability to diverse and noisy real-world conditions.

Table 3. Recap of performance metric comparison between CNN and MLP

Method	Metric	Experimental model							
		A	B	C	D	E	F	G	H
MLP	Accuracy	27.92	95.83	96.67	30.83	96.67	97.50	95.42	96.67
	Recall	26.65	95.91	96.44	29.67	96.44	97.64	95.05	96.65
	Precision	23.51	96.01	97.21	28.51	97.14	97.50	96.39	96.88
	F1-score	22.79	95.93	96.73	25.81	96.72	97.56	95.50	94.88
	Kappa	12.97	94.98	95.98	16.33	95.98	96.99	94.47	97.49
CNN	Accuracy	20.83	96.25	95.42	28.75	97.50	97.08	96.25	97.92
	Recall	22.26	96.07	95.55	29.69	97.50	96.95	96.21	97.90
	Precision	11.81	96.59	95.56	42.56	97.59	97.39	96.44	97.97
	F1-score	13.83	96.27	95.55	24.57	97.54	97.56	96.31	97.92
	Kappa	05.55	95.48	94.48	15.26	96.99	96.99	95.48	97.49

### 3.4. Discussion

The proposed classification model uses the CNN algorithm based on the model H experiment, applying the adding noise, time stretching, and pitch shifting techniques to add new synthetic data to the testing data. Then, to determine the characteristics of each audio sample, dialect language feature extraction is applied. The evaluation metrics indicate promising performance of the proposed model, with an accuracy of 97.92% for dialect classification on the current dataset. The recall score of 97.9% shows that the model effectively identifies most positive cases, while the precision score of 97.97% demonstrates its ability to minimize false positives. The F1-score, reaching 97.92%, suggests a balance between precision and recall, and a Cohen's Kappa score of 97.49% reflects substantial agreement between predictions and true labels. However, given the relatively limited dataset and scope of the study (six dialects and less than 2,000 samples), further evaluations on larger and more diverse datasets are needed to confirm the generalizability and robustness of the model. Additionally, challenges such as handling intonation variability and real-world noise conditions remain areas for future improvement.

The performance of the CNN model in this research was compared with that of previous research conducted by Ernawati and Riana [16]. Their study utilized an 1D CNN combined with MFCC for feature extraction on the Java-SED dataset, achieving an accuracy of 96.43%. In contrast, this research applied the same 1D CNN architecture and MFCC feature extraction to the Indonesian dialects dataset, which consists of



multiple regional dialects. As a result, the proposed method achieved an improved accuracy of 97.92%. This indicates that the architecture and methodology adapted from the previous study are highly effective when applied to a broader and more diverse dataset, highlighting the robustness and generalizability of the 1D CNN model in classifying regional dialects.

Further comparison with studies on dialect classification also reinforces these findings. A study by Amani *et al.* [24] applied x-vectors for feature extraction combined with support vector machine (SVM), achieving an accuracy of 87%. Although x-vectors are an advanced technique in speaker recognition, the reliance on SVM appears to limit classification performance compared to CNN. Another study by Ghafoor *et al.* used MFCC features with 1D CNN on a Kurdish dialect dataset, achieving an accuracy of 95.53%. Similarly, Karim *et al.* [25] adopted a more complex approach by combining MFCC, Mel spectrogram, poly-feature, and contrast as feature extraction techniques with 1D CNN, resulting in an accuracy of 96.5%.

Compared to these studies, the method proposed in this research achieved the highest accuracy at 97.92%. This suggests that the combination of 1D CNN and MFCC features is not only effective but also robust in handling variations within the Indonesian dialect dataset. Nonetheless, it is important to note that differences in dataset characteristics, size, and evaluation protocols across studies may influence the reported results. Therefore, further research is recommended to validate the generalizability and robustness of this approach across different datasets and real-world scenarios.

4. CONCLUSION

Based on the research conducted, several conclusions can be drawn as a foundation for recommendations and future research development. The use of data augmentation techniques such as adding noise, time stretching, and pitch shifting has significantly improved the performance of dialect classification for regional languages. The research results show that models implementing data augmentation techniques consistently recorded accuracy improvements on the current dataset, with CNN achieving an accuracy of 97.92% and MLP reaching 97.5%. This suggests that data augmentation can enhance model performance for dialect classification in controlled conditions. Feature extraction using MFCC also played a significant role in capturing the unique acoustic characteristics of each dialect. However, further research is needed to evaluate these techniques' effectiveness in more complex and diverse datasets, especially under real-world conditions. The CNN and MLP algorithms demonstrate strong performance in dialect classification, with CNN consistently showing better performance than MLP in terms of accuracy, precision, recall, F1-score, and kappa. However, there are still challenges regarding the model's ability to generalize to new data, especially when facing variations in dialect intonation and expression. Therefore, further research is needed to optimize the model to be more adaptable to a wider range of dialect variations.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to all respondents who willingly participated and contributed their voice recordings for the purposes of this study. Their support and cooperation were instrumental in the data collection process.

FUNDING INFORMATION

The authors state no funding involved.

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
Fahmi B. Marasabessy	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Dwiza Riana	✓	✓			✓	✓		✓		✓		✓		✓
Muji Ernawati	✓	✓	✓	✓	✓					✓			✓	

C : Conceptualization	I : Investigation	Vi : Visualization
M : Methodology	R : Resources	Su : Supervision
So : Software	D : Data Curation	P : Project administration
Va : Validation	O : Writing - Original Draft	Fu : Funding acquisition
Fo : Formal analysis	E : Writing - Review & Editing	



## CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.




## DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [DR], upon reasonable request.




## REFERENCES

- [1] K. Kasiyarno and S. Apriyanto, "The influence of globalisation on the shift in local language and cultural identity," *Journal Corner of Education, Linguistics, and Literature*, vol. 4, no. 3, pp. 372–383, Feb. 2025, doi: 10.54012/jcell.v4i3.435.
- [2] A. F. Aji *et al.*, "One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 7226–7249, doi: 10.18653/v1/2022.acl-long.500.
- [3] N. V. Wilmot, M. Vigier, and K. Humonen, "Language as a source of otherness," *International Journal of Cross Cultural Management*, vol. 24, no. 1, pp. 59–80, 2024, doi: 10.1177/14705958231216936.
- [4] C. Fotiou and K. K. Grohmann, "A small island with big differences? folk perceptions in the context of dialect levelling and Koineization," *Frontiers in Communication*, vol. 6, pp. 1–19, Jan. 2022, doi: 10.3389/fcomm.2021.770088.
- [5] C. Syam, S. Seli, and W. J. Abdu, "Dynamics of language interaction in multicultural urban communities: analysis of socio-cultural linguistic environment," *Society*, vol. 11, no. 2, pp. 575–588, Dec. 2023, doi: 10.33019/society.v11i2.628.
- [6] A. Musyaffa and L. S. Dewi, "An analytical study of language styles in different English dialects," *Jurnal Nakula : Pusat Ilmu Pendidikan, Bahasa dan Ilmu Sosial*, vol. 2, no. 5, pp. 222–231, Jul. 2024, doi: 10.61132/nakula.v2i5.1051.
- [7] A. Wang, "Speech recognition for different dialects and accents," *ITM Web of Conferences*, vol. 73, Feb. 2025, doi: 10.1051/itmconf/20257302011.
- [8] Q. Li, Q. Mai, M. Wang, and M. Ma, "Chinese dialect speech recognition: a comprehensive survey," *Artificial Intelligence Review*, vol. 57, no. 2, pp. 1–39, 2024, doi: 10.1007/s10462-023-10668-0.
- [9] K. J. Ghafoor, K. M. H. Rawf, A. O. Abdulrahman, and S. H. Taher, "Kurdish dialect recognition using 1D CNN," *Aro-The Scientific Journal of Koya University*, vol. 9, no. 2, pp. 10–14, Oct. 2021, doi: 10.14500/aro.10837.
- [10] D. E. -Grisales *et al.*, "Colombian dialect recognition based on information extracted from speech and text signals," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 556–563, doi: 10.1109/ASRU51503.2021.9687890.
- [11] B. Tawaqal and S. Suyanto, "Recognizing five major dialects in Indonesia based on MFCC and DRNN," *Journal of Physics: Conference Series*, vol. 1844, no. 1, Mar. 2021, doi: 10.1088/1742-6596/1844/1/012003.
- [12] K. Nugroho, E. Noersasongko, Purwanto, Muljono, and D. R. I. M. Setiadi, "Enhanced Indonesian ethnic speaker recognition using data augmentation deep neural network," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4375–4384, Jul. 2022, doi: 10.1016/j.jksuci.2021.04.002.
- [13] E. Shandy, A. H. Anshor, and D. Ardiatma, "Implementation of data mining for speech recognition classification of Sundanese dialect using KNN method with MFCC feature extraction," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 6, no. 3, pp. 1145–1158, 2024, doi: 10.47709/cnahpc.v6i3.4226.
- [14] I. G. A. G. A. Kadyanan *et al.*, "Balinese text-to-speech dataset as digital cultural heritage," *Data in Brief*, vol. 60, 2025, doi: 10.1016/j.dib.2025.111528.
- [15] D. Wang, S. Ye, X. Hu, S. Li, and X. Xu, "An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model," in *Interspeech 2021*, Aug. 2021, vol. 3, pp. 3266–3270, doi: 10.21437/Interspeech.2021-374.
- [16] M. Emawati and D. Riana, "Classification of human emotions based on Javanese speech using convolutional neural network and multilayer perceptron," *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, vol. 12, no. 1, pp. 101–112, Mar. 2024, doi: 10.52549/ijeel.v12i1.5343.
- [17] A.-O. Boudraa, K. Khaldi, T. Chonavel, M. T. H.-Alouane, and A. Komaty, "Audio coding via EMD," *Digital Signal Processing*, vol. 104, Sep. 2020, doi: 10.1016/j.dsp.2020.102770.
- [18] B. A. Alsaify, H. S. A. Arja, B. Y. Maayah, and M. M. Al-Taweel, "A dataset for voice-based human identity recognition," *Data in Brief*, vol. 42, Jun. 2022, doi: 10.1016/j.dib.2022.108070.
- [19] H. A. Alsayadi, A. A. Abdelhamid, I. Hegazy, B. Alotaibi, and Z. T. Fayed, "Deep investigation of the recent advances in dialectal Arabic speech recognition," *IEEE Access*, vol. 10, pp. 57063–57079, 2022, doi: 10.1109/ACCESS.2022.3177191.
- [20] I. G. B. A. P. Paramitha, H. B. Kusnawan, and M. Emawati, "Performance comparison of deep learning algorithm for speech emotion recognition," *Journal of Computer Science and Informatics Engineering (J-Cosine)*, vol. 6, no. 2, pp. 99–106, Dec. 2022, doi: 10.29303/jcosine.v6i2.443.
- [21] H. D. Saputra, A. I. E. Efendi, E. Rudini, D. Riana, and A. S. Hewiz, "Hepatitis prediction using K-NN, naive Bayes, support vector machine, multilayer perceptron, and random forest, gradient boosting, k-means," *Journal Medical Informatics Technology*, vol. 1, no. 4, pp. 96–100, Dec. 2023, doi: 10.37034/medinftech.v1i4.21.
- [22] F. Arifin, A. S. Priambodo, A. Nasuha, A. Winursito, and T. S. Gunawan, "Development of Javanese speech emotion database (Java-SED)," *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, vol. 10, no. 3, pp. 584–591, 2022, doi: 10.52549/ijeel.v10i3.3888.
- [23] C. Nugraha and S. Hadiani, "Glaucoma detection in fundus eye images using convolutional neural network method with visual geometric group 16 and residual network 50 architecture," *Journal Medical Informatics Technology*, vol. 1, no. 2, pp. 36–41, Jun. 2023, doi: 10.37034/medinftech.v1i2.7.
- [24] A. Amani, M. Mohammadamini, and H. Veisi, "Kurdish spoken dialect recognition using x-vector speaker embedding BT-speech and computer," in *23rd International Conference, SPECOM 2021*, 2021, pp. 50–57, doi: 10.1007/978-3-030-87802-3\_5.
- [25] S. H. T. Karim, K. J. Ghafoor, A. O. Abdulrahman, and K. M. H. Rawf, "A multi-feature fusion approach for dialect identification using 1D CNN," *International Journal on Informatics Visualization*, vol. 8, no. 3, pp. 1246–1252, 2024, doi: 10.62527/joiv.8.3.2146.




**BIOGRAPHIES OF AUTHORS**

**Fahmi B. Marasabessy, S.T., M.Kom.**    is a recent master's graduate from the Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, Indonesia, currently serving as the Chairperson of the Alumni Association of Informatics Engineering at Universitas Pancasila. Actively involved in several hobby-based communities, he is passionate about technology and community engagement. With a keen interest in informatics engineering, he is now focusing on advancing their expertise in areas such as software development, data analysis, and emerging technologies. He is open to collaborative opportunities in research and innovation. He can be contacted at email: 14220011@nusamandiri.ac.id.



**Prof. Dr. Ir. Dwiza Riana, S.Si, M.M., M.Kom., IPU, ASEAN.Eng.**    has been a permanent lecturer at the Faculty of Information Technology since 2003, and is currently a lecturer in the Doctoral Program in Informatics at Universitas Nusa Mandiri. 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. She can be contacted at email: dwiza@nusamandiri.ac.id.



**Muji Ernawati, M.Kom.**    is a lecturer at Universitas Nusa Mandiri, teaching in the Informatics Study Program at Faculty of Information Technology. Additionally, she is part of the Research and Community Service Institute (LPPM) team at Universitas Nusa Mandiri, specifically in the Intellectual Property Center and Information Systems unit. She completed her master's degree in Computer Science (M.Kom.) in 2023. She has published several research papers in national journals and conferences, focusing on machine learning and deep learning. Her research interests include machine learning, deep learning, and natural language processing. She can be contacted at email: muji.mei@nusamandiri.ac.id.