Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 2, May, Pp. 205-216

Maleo Emotion Audio Dataset Indonesia for Emotion Classification

Ardi Mardiana[1], Sri Mentari Widya Ningrum Permana[*1], Ii Supiandi[1], Ade Bastian[1], Eka Tresna Irawan[2]
University of Majalengka, Indonesia[1]
Mayar International Pte. Ltd, Singapore[2]

Article Info
Keywords:
Speech Emotion Recognition, Voice Emotion, Indonesian Language Dataset, CNN, Audio Augmentation

Article history:
Received: August 22, 2025
Accepted: December 01, 2025
Published: May 01, 2026

Cite:
A. Mardiana, S. M. W. N. Permana, Ii Sopiandi, Ade Bastian, and E. T. Irawan, "Maleo Emotion Audio Dataset Indonesia for Emotion Classification", *KINETIK*, vol. 11, no. 2, May. 2026.
https://doi.org/10.22219/kinetik.v11i2.2474

*Corresponding author.
Sri Mentari Widya Ningrum Permana
E-mail address:
agus_purwadi@polije.ac.id

Abstract
The limited availability of voice emotion corpora in Indonesian poses a challenge for the development of Speech Emotion Recognition (SER) systems, despite growing needs in sectors such as customer service and human-computer interaction. To address this, we developed the Maleo Emotion Audio Corpus, a collection of three-second audio clips with seven emotion labels (angry, neutral, disgusted, sad, happy, afraid, and surprised), sourced from YouTube. The audio data underwent preprocessing, feature extraction (MFCC, ZCR, energy, spectral roll-off, and spectral flux), and augmentation. The classification model was built using a 1D Convolutional Neural Network (CNN) architecture specifically adapted for the 3-second audio features, comprising four convolutional layers. Evaluation showed the model achieved 94.48% accuracy on the test data. The claim of balanced performance is supported by high F1-scores across all classes, ranging from 0.87 for 'sad' to 0.98 for 'neutral', indicating no single class dominated the results. These findings demonstrate that the developed corpus and model architecture have strong capability for recognizing emotions from Indonesian speech in a locally relevant context. Maleo Emotion collection is available at https://doi.org/10.57967/hf/6144.

1. Introduction

Emotions are states that comprehensively represents human feelings and thoughts and can be found in various aspects of daily life [1]. Extensive studies on emotions are currently being conducted across various sectors, as emotions are among the most fundamental elements of human beings [2]. Human-computer interaction technology has been advancing rapidly, such as speech recognition. One of the applications of speech recognition is the identification of human emotions [3]. Human interaction refers to the ways people interact with one another. However, human interaction does not always proceed smoothly due to various factors that can influence emotions [4]. Additionally, meeting the need for emotion recognition is crucial in human interaction, as it is beneficial for various applications, such as customer service, mental health, educational systems, and the development of simpler human-computer interfaces [5].

Emotion recognition technology through voice, known as Speech Emotion Recognition (SER), is a rapidly developing field crucial for improving human-computer interaction [6]. SER enables systems to understand and respond to users' emotional states in a more contextual and human-like manner, with applications ranging from customer service to mental health [7].

However, a significant challenge hinders the advancement of SER for the Indonesian language: the lack of a standardized and representative audio corpus [8]. Based on the background and research gaps described above, the issues that are the focus of this study are as follows: (1) To what extent can this dataset improve the accuracy of emotion classification systems compared to other datasets? (2) How can we determine which emotion categories should be included in the dataset? (3) How can we determine the reference standards that can be used in testing and evaluating Indonesian SER models? As a solution, this study developed the Maleo Emotion Audio Indonesia dataset, establishes seven main emotion categories, and uses CNN models and standard evaluation metrics to measure the performance of the emotion classification system.

The development of any effective SER system hinges on three core components: quality data collection, robust acoustic feature extraction, and an accurate classification algorithm [9].

This research addresses the data scarcity issue by first developing a new Indonesian emotion corpus. Our method then involves two main stages to classify emotions from this audio data:

1. Feature Extraction: We convert raw audio signals into meaningful representations by extracting five key acoustic features: Mel-Frequency Cepstral Coefficients (MFCC) [10], Zero Crossing Rate (ZCR) [11], energy [12], spectral roll-off [13], and spectral flux [14].

2. Classification: These extracted features are then used to train a Convolutional Neural Network (CNN) model to perform the final emotion classification [15].

The availability of a high-quality, representative corpus is a crucial component for developing an effective SER system. This challenge is particularly pronounced for the Indonesian language, where standard public corpora are still very limited. For context, much of the existing SER research has relied on well-established corpora in other languages. For example, the IEMOCAP corpus has been widely used for processing, which includes the stages of feature extraction, model training, and performance evaluation [16].

Therefore, this data collection is highly valuable for researchers in signal processing and artificial intelligence, aiding in the development of emotion classification models in artificial intelligence, machine learning, multimedia, and signal processing [1]. While many high-quality SER datasets exist, such as RAVDESS and IEMOCAP, they predominantly feature speakers of English or other European languages. Emotional expression—including tone, intonation, and prosody—is deeply influenced by cultural and linguistic context [18]. Consequently, models trained on these international corpora often fail to accurately capture the unique nuances of emotion in Indonesian speech, making them less effective for local applications.

To address this gap, it is necessary to use data that reflects natural Indonesian communication. YouTube stands out as a vast repository of authentic audio and video content for this purpose [17]. Although using data from such platforms presents challenges, such as potential sample limitations and subjectivity in annotation, it offers a way to build a more culturally relevant corpus. Furthermore, this approach is supported by Indonesian Law No. 28 of 2014 concerning Copyright, which permits the use of such works for research and educational purposes, provided the source is properly cited [19].

SER faces persistent challenges, namely the scarcity of labeled data, which is a major obstacle given the intensive data requirements of deep learning models. This scarcity often results in small, unbalanced datasets, which hinder model generalization [18]. Additionally, the lack of cultural and linguistic diversity in many corpora remains a barrier to developing locally relevant emotion recognition systems. In previous studies, commonly used speech emotion corpora include RAVDESS, Ryerson Multimedia Lab (RML), and EMO-DB. RAVDESS provides 1,440 English-language voice samples with various emotion labels such as neutral, sad, calm, afraid, happy, surprised, disgusted, and angry [19]. RML includes 684 samples in various languages such as English, Italian, Mandarin, Urdu, Punjabi, and Persian [20].

The initial steps in identifying the main problems were the lack of Indonesian-language datasets for SER, determining emotion categories that correspond to emotional expressions in Indonesian, and the lack of reference standards or models for Indonesian SER.

This study used a collection of 700 audio files collected independently from various videos available on the YouTube platform. These were then extracted into 3-second audio segments in WAV format. This dataset is named *Maleo Emotion*. The data is annotated with seven emotion labels: angry, surprised, happy, afraid, sad, disgusted, and neutral [21]. Each emotion label consists of 100 audio files.

The main objective is to support research in the field of audio processing and artificial intelligence, assist in the development of emotion-based technology, and improve understanding of emotional voice expressions in Indonesian. This collection is available at https://huggingface.co/datasets/maleo-ai/maleo-emotion. The use of a Convolutional Neural Network (CNN) combined with key acoustic features is a well-established and effective approach in SER research [22]. However, the application of this powerful method has not yet been widely tested in the context of the Indonesian language [23]

In this study, we apply this supervised learning algorithm to the newly developed *Maleo Emotion Corpus* [25]. The audio data have been labeled into seven distinct emotion classes: angry, surprised, happy, afraid, sad, disgusted, and neutral. The CNN is then trained to learn and differentiate these emotions by analyzing five core acoustic features: MFCC, Zero Crossing Rate (ZCR), spectral roll-off, spectral flux, and energy [24]. This corpus aims to serve as a comprehensive and inclusive resource for researchers and developers in creating accurate and relevant emotion classification models tailored to the Indonesian context.

2. Research Method

This study proposes the development of an Indonesian Speech Emotion Recognition (SER) model using the Maleo Emotion Audio Dataset. The research methodology consists of several stages: data collection, data preprocessing and augmentation, feature extraction, model development using a Convolutional Neural Network (CNN), and performance evaluation. All experiments were conducted using the Python programming language with the TensorFlow and Keras libraries on the Google Colaboratory platform. Figure 1 shows the research framework, illustrating the stages of the process from data collection to model evaluation.

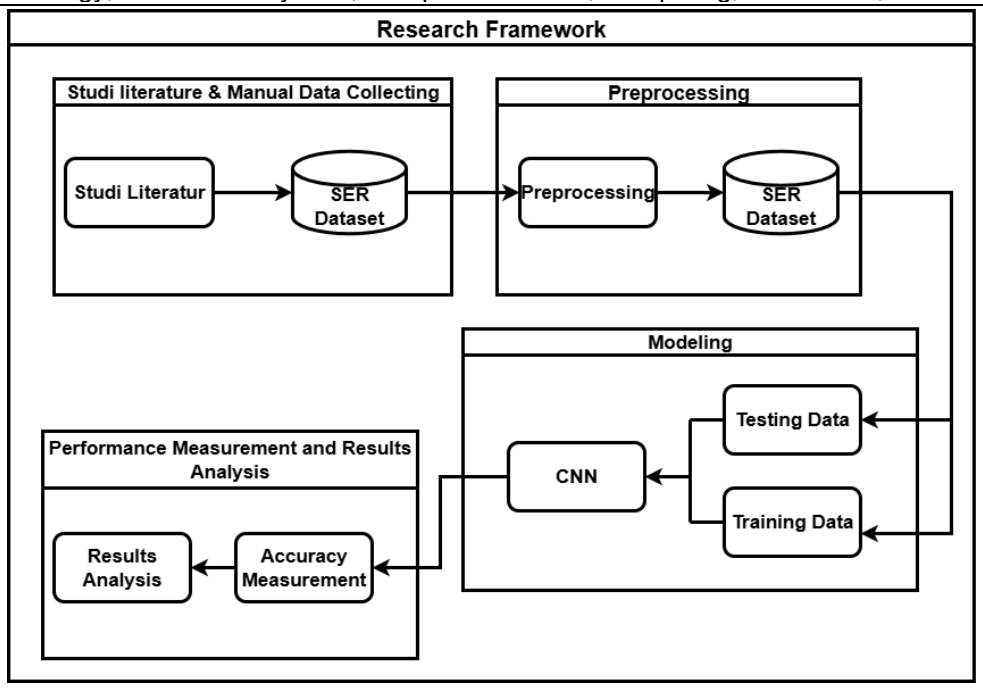

Figure 1. Research Framework

3. Data Collection

In this study, the audio dataset was manually collected from YouTube, consisting of 700 Indonesian voice samples representing seven distinct emotion categories: angry, neutral, disgusted, sad, happy, afraid, and surprised (see Table 1). Each sample is approximately 3 seconds long and is stored in .wav format. To maintain consistency, all audio files were standardized to mono with a sampling rate of 44,100 Hz and a resolution of 16-bit. This standardization ensured that the data had uniform characteristics during the feature extraction process. Data selection was performed manually to verify the accuracy of the emotions depicted in each clip. After verification, the data were grouped into folders based on their respective emotion labels. Finally, preprocessing steps such as background noise removal and audio normalization were applied to reduce noise and improve sound clarity.

Table 1. Audio Emotions

No	Emotion	Amount
1	Disgust	100
2	Anger	100
3	Neutral	100
4	Sadness	100
5	Happiness	100
6	Fear	100
7	Surprise	100

3.1 Waveform

A waveform is a graphical representation that illustrates changes in amplitude (loudness) of an audio signal over time. For a "surprised" emotion, the waveform typically displays a sharp spike in amplitude at the beginning, followed by a gradual decrease as the sound fades (see Figure 2). In this representation, the x-axis represents the time duration of the audio clip from 0 to 3 seconds, while the y-axis indicates the intensity of the sound. The sudden high peak at the start reflects the abrupt and strong vocal reaction typical of a surprised expression, followed by a continuous decrease in amplitude as the sound diminishes.

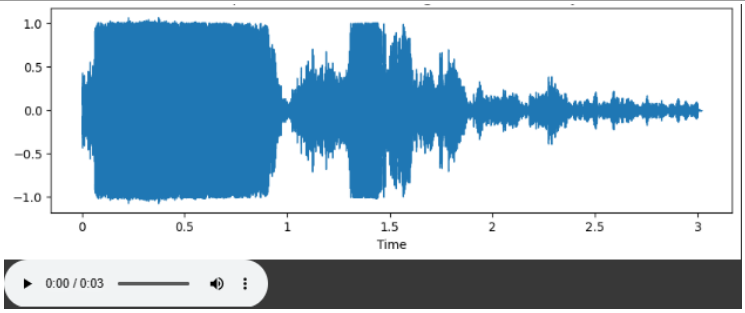

Figure 2. Waveform

3.2 Spectrogram

A spectrogram is a visualization that displays the frequency (pitch) of sound over time, where colors represent the amplitude or intensity at each frequency—with brighter colors, such as red, indicating higher intensity (see Figure 3). In this visualization, the x-axis represents the time duration of the audio from 0 to 3 seconds, while the y-axis displays the sound frequency range up to approximately 10,000 Hz. Although the audio was sampled at 44,100 Hz, the spectrogram is displayed up to 10,000 Hz because this range effectively captures the primary emotional information of the sound. The red and orange areas at the beginning of the recording indicate bursts of energy at low and mid frequencies, which are characteristic of the sudden vocalizations found in expressions of surprise. Over time, this energy spreads and gradually decreases.

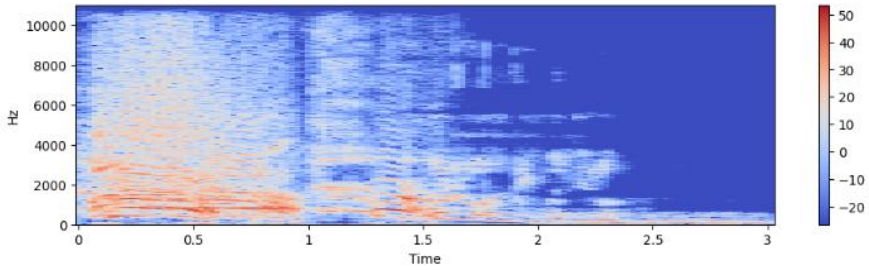

Figure 3. Spectrogram

Before the feature extraction process, the audio data undergoes an augmentation stage to increase both the variety and volume of the dataset. Following this, each audio file is verified to possess specific basic characteristics—such as file name, emotion label, duration, channels, sampling rate, and bit depth—to maintain strict data consistency (see Table 2).

Table 2. Property value

Property	Value
File Name	Surprised 1_mono.wav (example)
Emotion Label	Disgust
Duration	3 seconds
Channels	Mono
Sampling Rate	44,100 Hz
Bit Depth	32-bit

3.3 Data pre-processing

To prepare the audio data for model training, we implemented a two-stage preprocessing framework designed to ensure data quality and diversity. The first stage, data cleaning and standardization, focuses on creating a uniform set of audio files. The second stage, data augmentation, expands the dataset to prevent overfitting and improve the model's ability to generalize. The entire preprocessing workflow is illustrated in Figure 4.

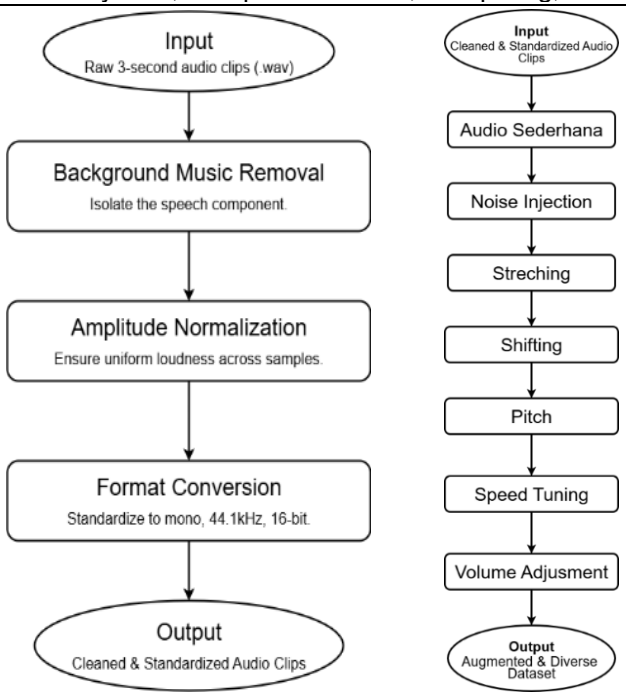

Figure 4. Preprocessing

3.4 Feature Extraction

In this study, we used five features for the extraction stage, including:

1. Zero Crossing Rate (ZCR)

The Zero-Crossing Rate (ZCR) within an audio frame measures the frequency of sign changes in the signal. Specifically, it represents how often the audio signal transitions from positive to negative, or vice versa, within a given audio segment. The ZCR for a specific frame is mathematically defined as Equation 1:

$$Z(i) = \frac{1}{2W_L}\sum_{n=1}^{W_L}|\mathrm{sgn}[x_i(n)] - \mathrm{sgn}[x_i(n-1)]| \quad (1)$$

2. Energy

Energy measures the total acoustic energy within an audio signal frame and is commonly associated with the intensity, or power, of the sound. Expressive emotions, such as anger or happiness, typically exhibit higher energy levels. The energy of an audio frame is calculated as Equation 2:

$$E(i) = \sum_{n=1}^{W_L}|x_i(n)|^2 \quad (2)$$

where $\mathrm{x_i(n)}$ represents the collection of audio samples from frame i, and W_L indicates the length (or duration) of that frame.

3. Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) are a widely utilized feature extraction technique in speech and audio signal processing. This technique adopts the Mel scale when calculating cepstral coefficients to better approximate human auditory perception. The primary purpose of this feature extraction is to capture the unique acoustic characteristics specific to different sounds or words, thereby enabling accurate differentiation between them. The MFCC calculation process is performed across the entire audio dataset in a single batch process and consists of the following sequential stages:

a) Pre-emphasis
The first step involves applying a pre-emphasis filter to amplify the high-frequency components of the audio signal, which naturally decay across the human vocal tract. This stabilizes the signal and is typically achieved using a first-order high-pass filter, as shown in Equation 3:

$$y(n) = x(n) - \alpha x(n-1) \tag{3}$$

where $x(n)$ is the original audio signal, $\mathrm{y(n)}$ is the pre-emphasized signal, and α is the pre-emphasis coefficient.

b) Framing
Because audio signals are non-stationary over long durations, the signal is divided into a series of short frames. Each frame is stored in an $\mathrm{M} \times \mathrm{W}$ matrix, where M represents the number of frames. Segmentation is performed using an overlapping method between frames to ensure that no critical acoustic information is missed at the boundaries. This process continues until the entire signal is covered.

c) Windowing
To minimize signal discontinuities at the beginning and end of each frame, a window function is applied. The Hamming window is commonly used to smooth the signal edges and reduce spectral leakage, as shown in Equation 4:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{4}$$

where N is the total number of samples within the frame.

d) Fast Fourier Transfrom (FFT)
The Fast Fourier Transform (FFT) is then applied to each windowed frame to convert the audio signal from the time domain into the frequency domain. This transformation reveals the magnitude spectrum of the frame, as shown in Equation 5:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}} \tag{5}$$

e) Mel Filterbank
The calculated frequency spectrum is then mapped onto the Mel scale using a set of triangular bandpass filters. The Mel scale is a non-linear mapping that mimics the human ear's varying sensitivity to different frequencies. The conversion from standard frequency (f in Hz) to the Mel scale (m) is mathematically formulated as Equation 6:

$$m = 1125\ln\left(1 + \frac{f}{700}\right) \tag{6}$$

f) Discrete Consine Trasfrom
Finally, in Equation 7, the Discrete Cosine Transform (DCT) is applied to the logarithm of the Mel filterbank energies. This step decorrelates the overlapping filterbank coefficients and compresses the data, yielding the final set of features used for model training, as shown below:

$$C(n) = \sum_{m=1}^{M} lo\,\mathrm{g}(E_m)\cos\left[n(m-0.5)\frac{\pi}{M}\right] \tag{7}$$

4. Spectrall Rollof

Spectral roll-off determines the frequency below which a specific percentage (typically 85% or 95%) of the total spectral energy is concentrated. This feature is commonly used to assess whether a sound's energy tends to be concentrated in lower or higher frequency bands. If the m-th Discrete Fourier Transform (DFT) coefficient represents the spectral roll-off value at the i-th frame, the relationship is defined by Equation 8:

$$\sum_{k=1}^{m} X_i(k) = C\sum_{k=1}^{W_L} X_i(k) \tag{8}$$

where $X_i(k)$ is the spectral magnitude at frequency bin k, W_L is the total number of bins in the frame, and C represents the chosen percentage threshold (e.g., 0.85).

5. Spectrall Flux

Spectral flux measures the rate of spectral change between two consecutive frames. It is calculated as the sum of the squared differences between the normalized magnitude spectra of the current frame and the previous frame. This feature effectively assesses how quickly the frequency profile of the signal changes over time. The spectral flux between frame i and frame $i-1$ is calculated as Equation 9:

$$Fl(i, i-1) = \sum_{k=1}^{W_L} \left(EN_i(k) - EN_{i-1}(k)\right)^2 \quad (9)$$

where $EN_i(k)$ represents the normalized spectral energy at bin k for frame i.

The values for all aforementioned features were extracted from the audio files using the Librosa library in Python. Following the extraction process, the resulting feature values were compiled into a structured dataset consisting of 4,900 rows and 25 columns. This consolidated dataset serves as the final, standardized input for the subsequent model training phase.

3.5 CNN Model Architecture

To classify the emotions based on the extracted acoustic features, a CNN model was employed. The architecture of the proposed CNN, illustrated in Figure 5, consists of several sequential layers designed to learn hierarchical representations of the audio data. The network begins with two convolutional layers utilizing a 3x3 kernel size—with 32 and 64 filters, respectively—to extract complex feature maps. These are followed by a pooling layer to reduce the spatial dimensions of the data and lower computational complexity. To mitigate the risk of overfitting during the training phase, a dropout layer is incorporated into the network. Finally, the architecture concludes with a dense (fully connected) layer that feeds into a softmax output layer, where the number of neurons corresponds directly to the number of emotion classes being predicted.

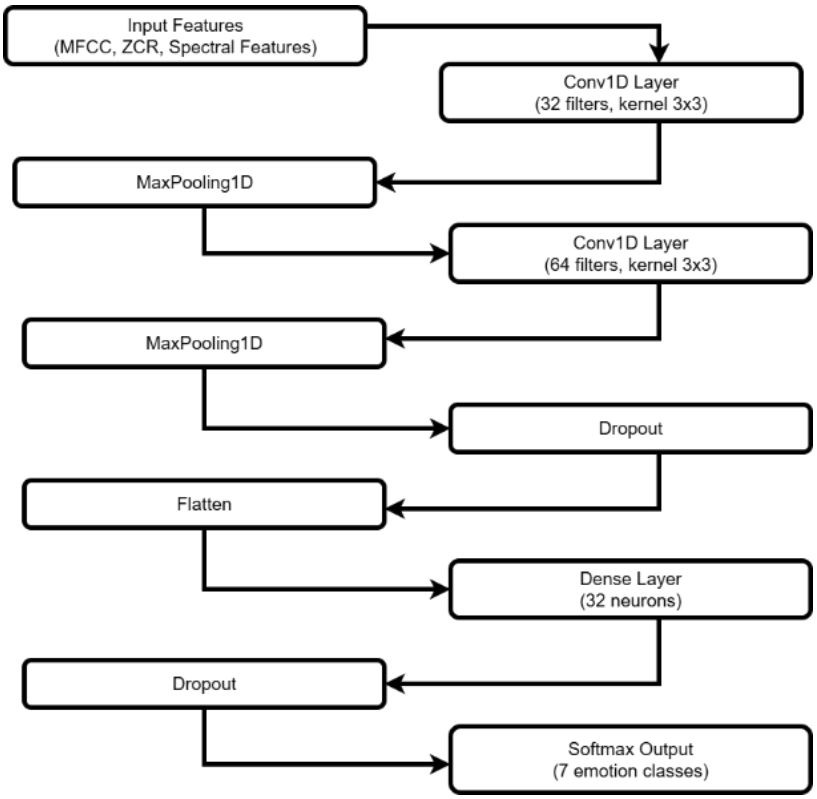

Figure 5. CNN model architecture used for Indonesian SER

3.6 Model Evaluation Analysis

To comprehensively assess the performance of the classification model, the following statistical evaluation metrics were utilized:

1. Accuracy

   Accuracy measures the overall proportion of correct predictions made by the model across all emotion classes. It is calculated as Equation 10, the ratio of correct predictions (n) to the total number of samples evaluated (N):

$$Accuracy = \frac{n}{N} \tag{10}$$

2. Precision

   Precision indicates the accuracy of the positive predictions for each specific class. It represents the ratio of true positive (TP) predictions to the total number of instances the model predicted as positive (which includes both true positives and false positives, FP), as shown in Equation 11:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

3. Recall

   Recall, also known as sensitivity, measures the model's ability to correctly identify all actual positive instances within a class. It is calculated as the ratio of true positive predictions to the total number of actual positive instances (which includes true positives and false negatives, FN), as shown in Equation 12:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

4. F1-Score

   The F1-Score is the harmonic mean of precision and recall. This metric is particularly useful because it provides a single, balanced performance score that accounts for both false positives and false negatives, as shown in Equation 13:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

4. Data Preparation

At this stage, the data was prepared, namely the dataset created from the YouTube platform, which consisted of 700 samples labeled with 7 emotions: happy, sad, angry, neutral, disgusted, afraid, and surprised. The dataset was stored in Google Drive.

4.1 Convolutional Neural Network Algorithm Model

The Convolutional Neural Network (CNN) model was implemented using the TensorFlow and Keras frameworks. To accelerate the training process and ensure optimal convergence, the network was optimized using the Adam optimizer. The complete architectural summary of the proposed CNN—detailing the sequential layers, their respective output shapes, and the number of trainable parameters—is presented in Table 3.

Table 3. Convolutional Neural Network Architecture Summary

Layer (Type)	Output Shape	Param #
conv1d	(None, 24, 256)	1,536
max_pooling1d	(None, 12, 256)	0
conv1d_1	(None, 12, 256)	327,936
max_pooling1d_1	(None, 6, 256)	0
conv1d_2	(None, 6, 128)	163,968
max_pooling1d_2	(None, 3, 128)	0
dropout	(None, 3, 128)	0
conv1d_3	(None, 3, 64)	41,024

max_pooling1d_3	(None, 2, 64)	0
flatten	(None, 128)	0
dense	(None, 32)	4,128
dropout_1	(None, 32)	0
dense_1	(None, 7)	231

4.2 Accuracy Data Results

After the training phase, the model was evaluated using a separate testing dataset to calculate its accuracy and assess its overall performance. The evaluation results demonstrated that the proposed CNN achieved a robust accuracy rate of 94.48% on the test data. This high accuracy indicates that the model is highly capable of correctly identifying emotions in unseen data that were not utilized during the training process. The performance of the model across all epochs is visualized in Figure 6, which displays both the loss and accuracy metrics.

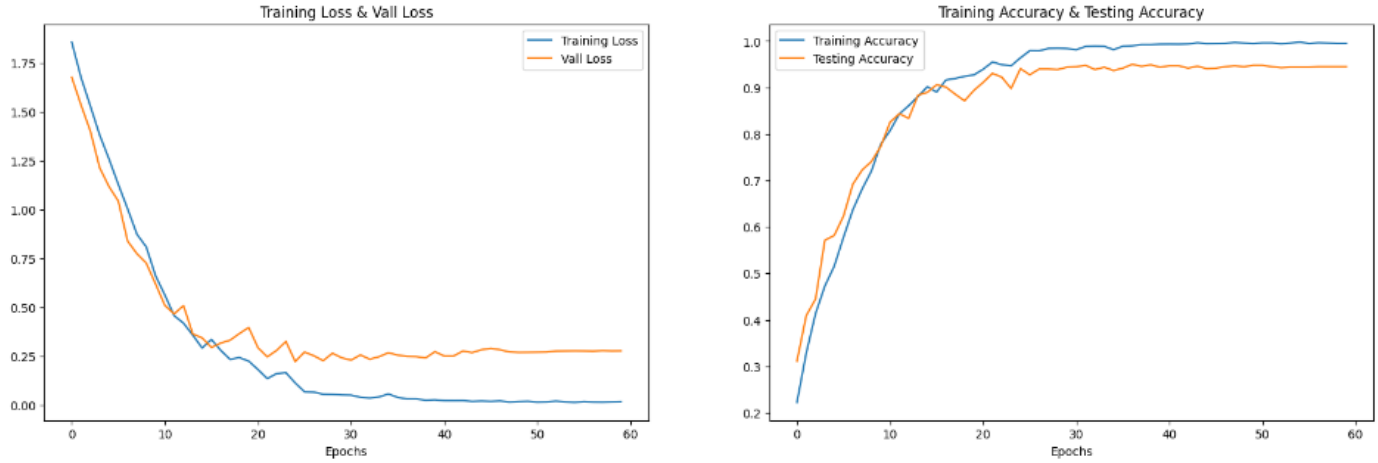

Figure 6. Training Loss & Val Loss and Training Accuracy & Testing Accuracy

4.2.1 Training and Validation Loss

The left panel of the figure illustrates the reduction in loss values for both the training and validation datasets over 60 epochs. Both the training loss (blue line) and validation loss (orange line) exhibit a sharp decline during the first 15 epochs. Following this initial phase, the training loss continues to decrease steadily toward zero, while the validation loss stabilizes at a low level without experiencing significant spikes. This behavior indicates that the model successfully learned the underlying patterns in the data without exhibiting severe signs of overfitting, maintaining a relatively small and stable gap between the two curves.

4.2.2 Training and Testing Accuracy

The right panel highlights the improvement in model accuracy throughout the training process. The accuracy on the training data (blue line) demonstrates a rapid increase, eventually approaching a near-perfect value. Concurrently, the accuracy on the testing data (orange line) displays a similarly strong and stable upward trend, settling at approximately 94.48%. The alignment and consistency between these two curves further confirm the model's capacity to generalize effectively to new data without suffering from performance degradation.

4.3 Confusion Matrix

To provide a detailed breakdown of the classification performance across the seven emotion categories (disgust, anger, neutral, sadness, happiness, fear, and surprise), a confusion matrix was generated (see Figure 7). This matrix illustrates the distribution of correct and incorrect predictions made by the model on the test dataset.

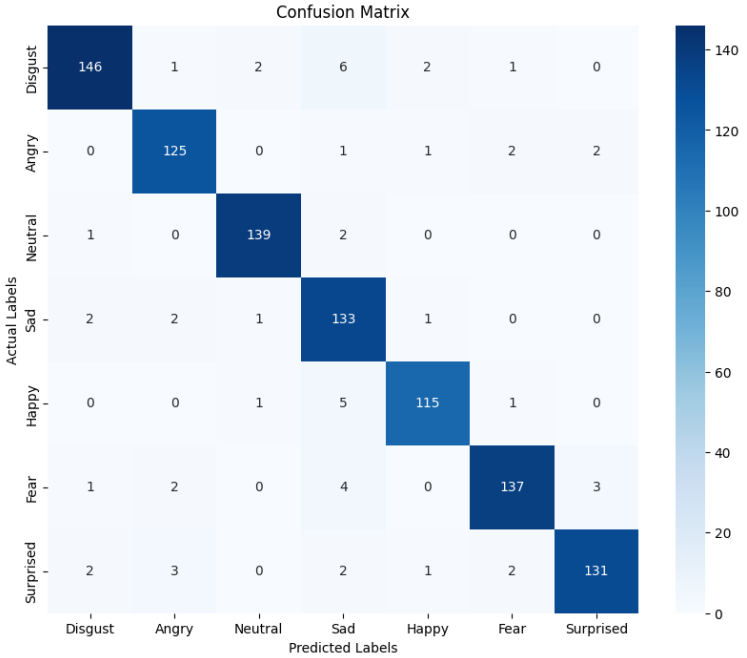

Figure 7. Confusion Matrix

The model demonstrated particularly strong performance in identifying disgust, neutral, and sad emotions. In the disgust category, 146 samples were correctly classified, with only minor misclassifications distributed across other labels (1 predicted as angry, 2 as neutral, 6 as sad, 2 as happy, and 1 as fearful). This indicates a high level of precision for the disgust class. Similarly, for the neutral emotion, the model recorded 139 correct predictions with minimal confusion, misclassifying only 1 instance as disgust and 2 as sad, demonstrating excellent recognition capabilities. Finally, for the sad emotion, 133 samples were correctly classified. The few errors in this category included instances incorrectly predicted as disgust (2 samples), angry (2 samples), neutral (1 sample), and happy (1 sample). Overall, these results highlight the model's high effectiveness and reliability in detecting these specific emotional states.

4.4 Classification Report

The developed voice emotion classification model demonstrated excellent performance, achieving a high overall accuracy rate. To comprehensively assess its classification effectiveness across the seven emotion categories (disgust, anger, neutral, sadness, happiness, fear, and surprise), the model was evaluated using standard statistical metrics, including precision, recall, and the F1-score, alongside the previously discussed confusion matrix. These metrics quantify how accurately and reliably the model identifies each specific emotion category. The detailed classification results are presented in Table 4.

Tabel 4. Classification Report

Emotion	Precision	Recall	F1-Score	Support
Disgust	0.96	0.92	0.94	158
Anger	0.94	0.95	0.95	131
Neutral	0.97	0.98	0.98	142
Sadness	0.87	0.96	0.91	139
Happiness	0.96	0.94	0.95	122
Fear	0.96	0.93	0.94	147
Surprise	0.96	0.93	0.95	141

In addition to the core performance metrics, Table 4 includes the *support* value, which denotes the actual number of occurrences for each emotion class within the testing dataset. Combining these granular metrics with the overall accuracy—which shows the model correctly classified 94.48% of the total test samples—provides a complete picture of the model's capabilities. Ultimately, these results demonstrate that the proposed model achieves highly accurate and well-balanced emotion recognition, confirming its reliability and effectiveness for voice-based emotion recognition tasks within the Indonesian language context.

5 Conclusion

The Maleo Emotion Audio Dataset Indonesia developed in this study can significantly improve the performance of audio-based emotion classification systems. With 700 audio samples, each lasting 3 seconds, the system built using the CNN algorithm achieved an accuracy of 94.48%, which is an improvement over previous studies reporting accuracy from 65–85%. This demonstrates that a locally relevant corpus, tailored to cultural and linguistic contexts, can enhance the effectiveness of SER systems. Emotion categories were determined based on literature and cultural context relevance in Indonesia. The selected emotions are anger, happiness, neutral, sadness, disgust, fear, and surprise, which are common expressions in Indonesian verbal communication. These categories reflect a sufficiently broad and relevant emotional range for use in various SER applications. Model evaluation was conducted using standard metrics such as accuracy, precision, recall, F1-score, and confusion matrix to assess the performance of each emotion category. Additionally, data augmentation techniques and balanced label distribution were employed to maintain the quality and validity of the dataset. These standards can serve as a baseline for testing Indonesian language-based SER models to ensure representative results that are comparable across studies.

References

[1] H. N. Zahra, M. O. Ibrohim, J. Fahmi, R. Adelia, F. A. Nur Febryanto, and O. Riandi, "Speech emotion recognition on indonesian youtube web series using deep learning approach," *2020 5th Int. Conf. Informatics Comput. ICIC 2020*, 2020. https://doi.org/10.1109/ICIC50835.2020.9288650

[2] A. Bustamin, A. M. Rizky, E. Warni, I. S. Areni, and I. Indrabayu, "IndoWaveSentiment: Indonesian Audio Dataset for Emotion Classification," *Mendeley Data*, vol. 1, 2024. https://doi.org/10.1016/j.dib.2024.111138

[3] D. Naresh Kumar, G. Deepak, and A. Santhanavijayan, "A Novel Semantic Approach for Intelligent Response Generation using Emotion Detection Incorporating NPMI Measure," *Procedia Comput. Sci.*, vol. 167, pp. 571–579, 2020. https://doi.org/10.1016/j.procs.2020.03.320

[4] D. Ardiyansyah and Jayanta, "Model Klasifikasi Emosi Berdasarkan Suara Manusia Dengan Metoode Multilater Perceptron," *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, no. April, pp. 689–702, 2021.

[5] T. B. Putri, S. Saidah, B. Hidayat, F. Qothrunnada, and D. Darwindra, "Deteksi Emosi Berdasarkan Sinyal Suara Manusia Menggunakan Discrete Wavelet Transform (DWT) Dengan Klasifikasi Support Vector Machine (SVM)," *J. Ilmu Komput. dan Inform.*, vol. 3, no. 1, pp. 1–10, 2023. https://doi.org/10.54082/jiki.45

[6] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 4040–4054, 2020. https://doi.org/10.18653/v1/2020.acl-main.372

[7] F. Kasyidi, R. Ilyas, and N. M. Annisa, "Peningkatan Kemampuan Pengenalan Emosi Melalui Suara dalam Bahasa Indonesia," *MIND J.*, vol. 6, no. 2, pp. 194–204, 2021. https://doi.org/10.26760/mindjournal.v6i2.194-204

[8] S. K. Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, "Speech based Emotion Recognition using Machine Learning," *2021 IEEE Mysore Sub Sect. Int. Conf. MysuruCon 2021*, no. Iccmc, pp. 613–617, 2019. https://doi.org/10.1109/ICCMC.2019.8819858

[9] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, 2020. https://doi.org/10.1016/j.specom.2019.12.001

[10] O. U. Kumala and A. Zahra, "Indonesian Speech Emotion Recognition using Cross-Corpus Method with the Combination of MFCC and Teager Energy Features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, pp. 163–168, 2021. https://doi.org/10.14569/IJACSA.2021.0120422

[11] F. R. K. Andre Julio Sumurung Marbun, Heriyanto, "Implementation of Mel-Frequency Cepstral Coefficient as Feature Extraction Method On Speech Audio Data," vol. 21, no. 3, pp. 260–270, 2024. https://doi.org/10.31315/telematika.v21i3.12339

[12] A. G. Jondya and B. H. Iswanto, "Analisis dan Seleksi Fitur Audio pada Musik Tradisional Indonesia," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 77, 2018. https://doi.org/10.24014/coreit.v4i2.6506

[13] S. Helmiyah, A. Fadlil, and A. Yudhana, "Pengenalan Pola Emosi Manusia Berdasarkan Ucapan Menggunakan Ekstraksi Fitur Mel-Frequency Cepstral Coefficients (MFCC)," *CogITo Smart J.*, vol. 4, no. 2, pp. 372–381, 2019. https://doi.org/10.31154/cogito.v4i2.129.372-381

[14] M. M. Billah, M. L. Sarker, and M. A. H. Akhand, "KBES: A dataset for realistic Bangla speech emotion recognition with intensity level," *Data Br.*, vol. 51, p. 109741, 2023. https://doi.org/10.1016/j.dib.2023.109741

[15] V. Sareen and K. R. Seeja, "Speech Emotion Recognition using Mel Spectrogram and Convolutional Neural Networks (CNN)," *Procedia Comput. Sci.*, vol. 258, pp. 3693–3702, 2025. https://doi.org/10.1016/j.procs.2025.04.624

[16] R. Y. Rumagit, G. Alexander, and I. F. Saputra, "Model Comparison in Speech Emotion Recognition for Indonesian Language," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 789–797, 2021. https://doi.org/10.1016/j.procs.2021.01.098

[17] F. Fahmi, M. A. Jiwanggi, and M. Adriani, "Speech-Emotion Detection in an {I}ndonesian Movie," *Proc. 1st Jt. Work. Spok. Lang. Technol. Under-resourced Lang. Collab. Comput. Under-Resourced Lang.*, no. May, pp. 185–193, 2020.

[18] G. Liu, S. Cai, and C. Wang, "Speech emotion recognition based on emotion perception," *Eurasip J. Audio, Speech, Music Process.*, vol. 2023, no. 1, 2023. https://doi.org/10.1186/s13636-023-00289-4

[19] I. Dewa Agung Adwitya Prawangsa and A. Eka Karyawati, "Penerapan Metode MFCC dan LSTM untuk Speech Emotion Recognition," *J. Elektron. Ilmu Komput. Udayana*, vol. 12, no. 4, pp. 2654–5101, 2024.

[20] A. T. Puspasari and A. Sardjono, "Pembatasan Hak Cipta Terkait Remix Lagu Berdasarkan Doktrin Fair Use Dan Undang- Undang Nomor 28 Tahun 2014 Tentang Hak Cipta," *J. Huk. Pembang.*, vol. 2, no. 2, 2023. https://doi.org/10.21143/telj.vol2.no2.1040

[21] S. Kakuba and D. S. Han, "Addressing data scarcity in speech emotion recognition: A comprehensive review," *ICT Express*, vol. 11, no. 1, pp. 110–123, 2025. https://doi.org/10.1016/j.icte.2024.11.003

[22] F. Jonatan Tanudjaja, E. Y. Puspaningrum, and V. Via, "Klasifikasi Jenis Emosi Melalui Ucapan Menggunakan Metode Convolutional Neural Network Type Of Emotions Classification Based On Speech Using Convolutional Neural Network Method," *Online) Teknol. J. Ilm. Sist. Inf.*, vol. 13, no. 2, pp. 1–11, 2023.

[23] A. Slimi, N. Haffar, M. Zrigui, and H. Nicolas, "Multiple Models Fusion for Multi-label Classification in Speech Emotion Recognition Systems," *Procedia Comput. Sci.*, vol. 207, no. Kes, pp. 2875–2882, 2022. https://doi.org/10.1016/j.procs.2022.09.345

[24] Riccosan, K. E. Saputra, G. D. Pratama, and A. Chowanda, "Emotion dataset from Indonesian public opinion," *Data Br.*, vol. 43, no. June 2024, 2022. https://doi.org/10.1016/j.dib.2022.108465

[25] Rini Andriani, Rizki Risdah Sitorus, Samuel Anaya Putra Zai, and Yesika Syalomi Pasaribu, "Penggunaan Algoritma CNN untuk Mengidentifikasi Jenis Anjing Menggunakan Metode Supervised Learning," *Mutiara J. Penelit. dan Karya Ilm.*, vol. 1, no. 6, pp. 393–403, 2023. https://doi.org/10.59059/mutiara.v1i6.741