JIKO (Jurnal Informatika dan Komputer) Accredited KEMDIKTISAINTEK, No.0173/C3/DT.05.00/2025
Vol. 8, No. 3, Desember 2025, pp. 244-251 p-ISSN: 2614-8897
DOI: 10.33387/jiko.v8i3.10820 e-ISSN: 2656-1948

IMPROVING INDONESIAN SPEECH EMOTION CLASSIFICATION USING MFCC AND BILSTM WITH AUDIO AUGMENTATION

Muhammad Septiyanto[1*], Eko Budi Susanto[2], Devi Sugianti[3]

[1,2,3] Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Institut Widya Pratama, Pekalongan, 51116, Indonesia
Email: [*1]msyant990@gmail.com, [2]ekobudi@stmik-wp.co.id, [3]devisugianti.iwp@gmail.com

(Received: 8 October 2025, Revised: 20 October 2025, Accepted: 5 November 2025)

Abstract

Emotion classification from speech has become an important technology in the modern artificial intelligence era. However, research for the Indonesian language is still limited, with existing methods predominantly relying on conventional machine learning approaches that achieve a maximum accuracy of only 90%. These traditional methods face challenges in capturing complex temporal dependencies and bidirectional contextual patterns inherent in emotional speech, particularly for Indonesian prosodic characteristics. To address this limitation, this study uses a combination of Mel-Frequency Cepstral Coefficients (MFCC) feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) model with audio augmentation techniques for Indonesian speech emotion classification. The IndoWaveSentiment dataset contains 300 audio recordings from 10 respondents with five emotion classes: neutral, happy, surprised, disgusted, and disappointed. Audio augmentation techniques with a 2:1 ratio using five methods generated 900 samples. MFCC feature extraction produced 40 coefficients that were processed using BiLSTM architecture with two bidirectional layers (256 and 128 units). The model was trained using Adam optimizer with early stopping. Research results show the highest accuracy of 93.33% with precision of 93.7%, recall of 93.3%, and F1-score of 93.3%. The "surprised" emotion achieved perfect performance (100%), while "happy" had the lowest accuracy (88.89%). This result surpasses previous benchmarks on the same dataset, which utilized Random Forest (90%) and Gradient Boosting (85%). This study demonstrates the effectiveness of combining MFCC, BiLSTM, and audio augmentation in capturing Indonesian speech emotion characteristics for the development of voice-based emotion recognition systems.

Keywords: *Emotion Classification, MFCC, BiLSTM, Audio Augmentation, Deep Learning*

**Corresponding Author: Muhammad Septiyanto*

1. INTRODUCTION

In the era of rapid artificial intelligence (AI) advancement, Speech Emotion Recognition (SER) has emerged as a critical technology enabling computers to interpret and respond to human emotions through vocal characteristic analysis [1]. The widespread adoption of AI-powered voice assistants such as Siri, Alexa, and Google Assistant demonstrates the increasing capability of intelligent systems to understand the contextual nuances of human communication [2]. Indonesia's position as the fourth most AI enthusiastic country globally presents a significant opportunity for developing SER technology tailored to Indonesian vocal characteristics [3]. However, the development of emotion recognition systems specifically designed for the Indonesian language remains limited [4], underscoring the urgent need for research that considers linguistic and cultural factors influenced by unique prosodic characteristics [5].

Recent studies have demonstrated progress in Indonesian speech emotion recognition. Bustamin et al. (2024) introduced the IndoWaveSentiment dataset containing 300 Indonesian emotional audio recordings across five classes: neutral, happy, surprised, disgusted, and disappointed, which were validated through manual annotations and questionnaires [6]. Building upon this dataset, Majiid et al. (2025) conducted comparative studies using conventional machine learning methods, with Random Forest achieving 90% accuracy through the combination of spectral contrast and MFCC features [7]. Prawangsa and Karyawati (2024) applied MFCC and LSTM

techniques on the TESS dataset, achieving a validation accuracy of 72.32% [8]. In related deep learning applications for Indonesian language processing, Mariyanto and Pardede (2023) demonstrated the effectiveness of LSTM networks in sentiment analysis tasks, achieving 77.96% accuracy on Indonesian text data, highlighting the potential of LSTM architectures for capturing sequential patterns in Indonesian language contexts [9]. In related acoustic classification domains, Zhang et al. (2023) achieved 93.81% accuracy for urban forest sound classification using deep learning [10], Dias et al. (2025) obtained up to 91% accuracy by integrating spectrograms and acoustic indices for bird and frog sound classification [11], and Kumar et al. (2024) demonstrated that audio augmentation significantly enhanced model robustness and accuracy in bird sound classification [12]. Similar advancements in music classification [13] emphasize the potential of combining spectral features, augmentation, and deep architectures for improved performance.

Despite these advancements, existing state-of-the-art methods for Indonesian speech emotion recognition predominantly rely on conventional machine learning approaches, with the highest reported accuracy reaching 90% using Random Forest [7]. These methods are limited in capturing complex temporal dependencies and bidirectional contextual patterns inherent in emotional speech. Moreover, existing research has not yet optimized deep learning architectures that reflect the prosodic and phonological characteristics unique to the Indonesian language particularly variations in intonation, duration, and fundamental frequency influenced by diverse linguistic and cultural backgrounds [5]. The absence of studies integrating bidirectional temporal processing and Indonesian specific audio augmentation strategies represents a substantial gap constraining the development of robust SER systems for Indonesian contexts.

To address this gap, this research proposes a novel combination of MFCC feature extraction and BiLSTM architecture enhanced with Indonesian-adapted audio augmentation for speech emotion classification. The novelty of this study is threefold: (1) optimization of MFCC parameters calibrated to the IndoWaveSentiment dataset characteristics; (2) implementation of five audio augmentation techniques designed to capture Indonesian prosodic variability; and (3) application of BiLSTM architecture to learn bidirectional temporal dependencies in Indonesian emotional speech signals. MFCC was chosen for its proven capability in extracting spectral features through mel-scale transformation that reflects human auditory perception [13], audio augmentation techniques were incorporated to enhance model robustness against vocal variations [12], and the BiLSTM architecture was employed for its superior ability to learn temporal patterns from both forward and backward directions [14]. The primary objective of this research is to develop an accurate and robust Indonesian speech emotion classification model that surpasses existing performance benchmarks while maintaining strong generalization across diverse speakers and emotional intensities.

2. RESEARCH METHOD

The research methodology is depicted in Figure 1, which provides guidelines on the flow of research stages. It consists of data collection, preprocessing, augmentation, feature extraction, modeling and evaluation.

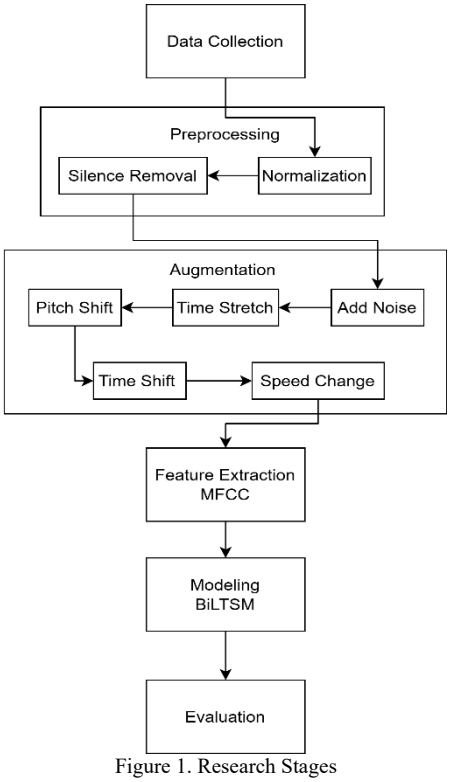

Figure 1. Research Stages

The research was conducted using a deep learning model, namely BiLSTM, for the purpose of classifying the emotions of Indonesian voices. The following is a series of research.

2.1 Data Collection

The dataset used in this study was obtained from IndoWaveSentiment, an Indonesian-language emotion audio dataset developed by Bustamin et al. (2024) in their previous research. (2024) [6]. This dataset contains voice recordings from 10 respondents (5 male and 5 female) who are radio announcers, singers, and professional voice actors. Each respondent uttered the sentence “The quality of this phone is pretty good” with five emotion variations, namely neutral, happy, surprised, disgusted, and disappointed.

Each emotion was recorded with two levels of intensity and repeated three times. Thus, each respondent produced 30 audio recordings, and a total of 300 audio files in .wav format were collected. These audio files are labeled according to a naming rule

consisting of actor identity, emotion class, intensity level, and repetition order. The dataset can be seen in table 1.

Table 1. Voice Emotion Dataset

Emotion	Intensity	Total Audio Male	Total Audio Female	Total
Neutral	Normal	15	15	30
Neutral	Strong	15	15	30
Happy	Normal	15	15	30
Happy	Strong	15	15	30
Surprised	Normal	15	15	30
Surprised	Strong	15	15	30
Disgusted	Normal	15	15	30
Disgusted	Strong	15	15	30
Disappointed	Normal	15	15	30
Disappointed	Strong	15	15	30
	Grand Total	150	150	300

2.2 Data Preprocessing

The preprocessing stage prepares the raw audio data into a format suitable for further processing by performing several sequential steps. The collected IndoWaveSentiment Indonesian emotion audio dataset in WAV format was loaded at a sample rate of 16 KHz. The preprocessing process includes audio normalization to ensure amplitudes fall within the range [-1, 1], followed by the removal of silent segments or noise with the parameter top_db=20. The dataset consists of 300 original audio files with a balanced distribution.

2.3 Augmentation

Standard data augmentation techniques were applied to increase dataset diversity and reduce overfitting, using an augmentation ratio of 2:1, resulting in a total of 900 samples from the original 300 samples. Five augmentation techniques were randomly implemented on each audio sample as follows:

1. Add Noise
   The addition of random Gaussian noise with a noise factor of 0.005 to simulate varying environmental conditions. This technique follows an approach that has proven effective in previous research to improve the robustness of the model [14].
2. Time Stretch
   Modifies the audio duration without affecting pitch, using a factor ranging from 0.8 to 1.2, to handle temporal variations in emotional expressions.
3. Pitch Shift
   Shifts the voice pitch by -3 to +3 semitones without changing the duration, to simulate individual differences.
4. Time Shift
   Circular temporal shift with a maximum of 20% of the audio length.
5. Speed Change
   Changing the speed through resampling by a factor of 0.9-1.1, which allows the model to recognize emotions at various speech speeds.

Each augmentation result is renormalized to maintain data consistency. This approach is inspired by audio augmentation strategies that have been shown to improve model robustness in speech classification. Previous research has shown that various augmentation techniques can improve accuracy up to 89.33% on the RAVDESS dataset [15]. In addition, the augmentation process can increase the dataset size to almost three times that of the original dataset [14], which contributes significantly to the improvement of the model performance.

2.4 Feature Extraction

Feature extraction using Mel-Frequency Cepstral Coefficients (MFCC) plays a crucial role in Indonesian speech emotion recognition, enabling the system to identify unique characteristics of emotional expressions. The MFCC algorithm transforms the representation of audio signals from the time domain to the frequency domain, mimicking human auditory perception to capture the essential elements of emotional sounds, in accordance with the approach used in speech analysis.

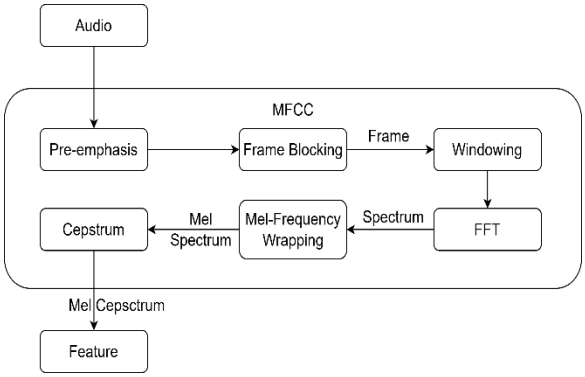

Figure 2. Mel-Frequency Cepstral Coefficient Process

1. Pre-emphasis is applied to audio data to increase the amplitude of high frequencies, reduce noise, and enhance the spectral shape of the signal. This process results in a spectrum with higher values at low frequencies that decrease gradually above 2000 Hz, in accordance with sound signal processing techniques.
2. Frame Blocking: The audio signal is divided into short frames (20-40 ms) with an overlap of about 50%, transforming the dynamic signal into a relatively stable sequence for frequency analysis using the Fourier transform. This process is applied to the entire duration of the original recording to ensure temporal consistency.
3. Windowing: Each frame is processed using a windowing function, such as a Hamming window, to minimize the tran sien effect at the edges of the signal, resulting in a smoother signal for spectral analysis.

4. Fast Fourier Transform (FFT) : The FFT converts a signal from the time domain to the frequency domain to generate a power spectrum [12]. This process enables the analysis of the frequency components of each emotion audio frame.
5. Mel-Frequency Wrapping : The power spectrum of each frame is filtered using a Mel-scale filter bank with 20-40 triangular filters. These filters mimic human auditory perception which is linear below 1 kHz and logarithmic above, resulting in filter energies that represent sound characteristics [13].
6. Logarithms (Log) are applied to the Mel filter energy to compress the dynamic range, resulting in a log Mel-spectrum that aligns with the perceived intensity of the human voice.
7. Discrete Cosine Transform (DCT) converts the log Mel-spectrum to the cepstral domain, resulting in 40 MFCC coefficients that provide a comprehensive feature representation for voice emotion analysis.

The implementation is automated using an audio processing library, extracting 40 MFCC coefficients. The extracted results are temporally averaged to produce a 40-dimensional feature vector based on the original duration of the recording.

2.5 Modeling

Bidirectional Long Short Term Memory (BiLSTM) is a variant of LSTM that processes sequences of data forward and backward to capture temporal context from both directions. BiLSTM is a development of Long Short Term Memory (LSTM) by connecting two hidden input layers in BiLSTM, namely forward inputs used to represent previous information and backward inputs used to represent later information [16]. This model is very suitable for analyzing the emotions of Indonesian voices based on MFCC features extracted with additional augmentation.

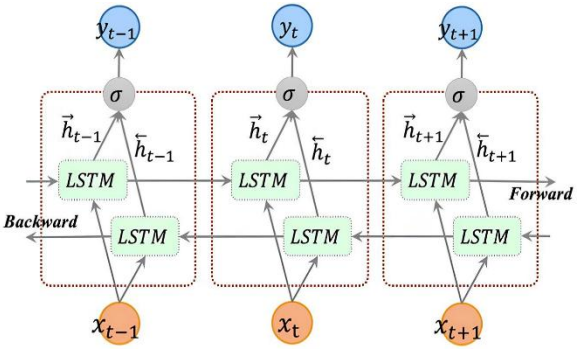

Figure 3. Architecture of BiLSTM Model

Figure 3 depicts the implemented BiLSTM architecture, consisting of two bidirectional LSTM layers followed by dropout and batch normalization, respectively. The hidden layer uses ReLU activation, while the output layer utilizes softmax activation for multi-class classification.

The selection of BiLSTM is based on its ability to handle complex patterns in emotion audio signals, where key information may be scattered across different parts of the recording. The bidirectional approach allows the model to learn the context from both directions simultaneously, improving the detectability of emotional features that may be missed with unidirectional models.

Table 2. Details Of BiLSTM Model

Layer Type	Output Shape	Parameters
Bidirectional LSTM	(None, 40, 256)	133,120
Batch Normalization	(None, 40, 256)	1,024
Bidirectional LSTM	(None, 128)	164,352
Batch Normalization	(None, 128)	512
Dense (ReLU)	(None, 128)	16,512
Dropout (default)	(None, 128)	0
Batch Normalization	(None, 128)	512
Dense (ReLU)	(None, 64)	8,256
Dropout (default)	(None, 64)	0
Batch Normalization	(None, 64)	256
Dense (Softmax)	(None, 5)	325
	Trainable parameters	323717
	Non-trainable parameters	1152
	Total Parameters	324869

Based on Table 2, a two-layer BiLSTM configuration with 256 and 128 units was chosen as it resulted in a total of 324,869 parameters that were optimal for the Indonesian voice emotion dataset. The use of default dropout maintains the balance between regularization and learning, while batch normalization after each layer ensures training stability with only 1,152 non-trainable parameters.

The dense layer with ReLU activation provides the necessary non-linearity for complex emotion classification (with 16,512 and 8,256 parameters), and the softmax output layer with 325 parameters is suitable for classifying five emotion classes. This approach is consistent with the utilization of BiLSTM to capture bidirectional context relationships in voice data.

Table 3. Hyperparameter Tuning BiLSTM

Parameter	Value
Batch size	32
Epochs	150
Loss Function	Sparse Categorical Crossentropy
Optimizer	Adam
Learning Rate	0.001 (adapted with ReduceLROnPlateau, min=0.0001, factor=0.5)
Early Stopping	Monitor='val_loss', patience=20, restore_best_weights=True
Model Checkpoint	Monitor='val_loss', save_best_only=True
Input shape	(40,1) – From X_train.shape
BiLSTM layers	2 Bidirectional LSTM layers (256 , 128 units)
Dense Layers	2 Dense layers (128, 64 units) + output layer (5 classes)
Activation Funtions	ReLU (hidden), Softmax (output)
Dropout rate	Default
Batch normalization	Applied after each layer

Based on Table 3, a batch size of 32 was chosen to balance computational efficiency and training stability. An initial learning rate of 0.001 with Adam optimizer was chosen as it provides stable convergence for sequential data. Early stopping with patience was applied as it prevents overfitting while allowing enough time for optimal convergence. The sparse categorical crossentropy loss function was chosen as it is compatible with integer labels for the five emotion classes, while dropout and batch normalization effectively cope with high variation in the audio data. An epoch setting of up to 150 was balanced with early stopping to avoid overfitting.

2.6 Confusion Matrix

To assess the effectiveness of the Indonesian voice emotion classification method using MFCC and BiLSTM, the classification results are analyzed to measure the success rate. Confusion matrix is the main evaluation tool in measuring model performance in multi-class classification, especially in voice emotion recognition that has significant acoustic variations. Confusion Matrix provides more specific information on how different genres can be identified using classification techniques [17]. The analysis involves calculating metrics such as accuracy, precision, recall, and F1-score obtained from the components of the matrix.

In voice emotion recognition research, every machine learning problem requires a customized set of metrics to accurately evaluate its performance [18]. The Confusion Matrix presents the relationship between model predictions and actual labels in a tabular format, with the vertical axis showing the original emotion categories and the horizontal axis showing the predicted categories from the BiLSTM model. Each cell in the matrix reflects the number of samples of a particular emotion classified to a particular prediction class.

1. Accuracy describes the overall ability of the model to correctly predict voice emotions from the total audio samples tested. Formula :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} .100 \quad (1)$$

2. Precision indicates the level of accuracy of the voice emotion prediction generated by the BiLSTM model compared to the targeted emotion. Formula :

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3. Recall measures the model's ability to detect all audio samples that truly represent a particular emotion. Formula :

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

4. F1-Score combines precision and recall to provide an optimal balance in evaluating the performance of voice emotion classification. Formula :

$$F1\ Score = 2.\frac{Recall.Precision}{Recall+Precision} \quad (4)$$

TP represents correctly predicted positive samples, TN represents correctly predicted negative samples, FP represents incorrectly predicted positive samples, and FN represents incorrectly predicted negative samples.

3. RESULT AND DISCUSSION

3.1 Data Collection

This research utilizes a public dataset obtained from IndoWaveSentiment, an Indonesian-language emotion audio dataset developed by previous research Bustamin et al. (2024) [6] which can be seen in table 1, with audio data obtained from the Mendeley Data platform under the title IndoWaveSentiment. Each emotion has unique vocal characteristics, so the selection of this dataset aims to test the model's ability to distinguish various voice emotions.

3.2 Preprocessing

This stage begins with loading audio files at a sampling frequency of 16 kHz to ensure processing consistency. Amplitude normalization was applied to the range [-1, 1] to standardize the volume, followed by removal of silent parts using the parameter top_db=20 to reduce background noise while preserving important vocal characteristics.

3.3 Data Augmentation

Five standard augmentation techniques were randomly implemented in a 2:1 ratio to increase the variability of the dataset from 300 to 900 samples. The applied techniques include Gaussian noise addition (factor of 0.005), time stretching with a factor of 0.8-1.2, pitch shifting ±3 semitones, circular time shift of maximum 20%, and speed change with a factor of 0.9-1.1. Each augmentation result is renormalized to maintain data consistency.

3.4 Feature Extraction

Feature extraction uses 40 MFCC coefficients implemented through mel-scale transform to capture the spectral characteristics of voice emotions. The process starts with pre-emphasis for spectral enhancement, frame blocking with a duration of 20-40 ms, windowing using Hamming function, FFT transformation to frequency domain, filtering with Mel-scale filter bank, logarithm application, and finally DCT to generate MFCC coefficients. The extraction results are temporally averaged resulting in a 40-dimensional feature vector.

Table 4. MFCC Feature Statistics Between Emotion Classes

Class	Min	Max	Mean ± SD	CV(%)
Neutral	-183.8	87.4	-48.2 ± 45.3	94.0
Happy	-174.2	77.3	-48.4 ± 42.2	87.2
Surprised	-166.4	70.8	-47.8 ± 39.7	83.1
Disgusted	-218.2	76.4	-70.9 ± 48.7	68.7
Disappointed	-222.3	81.3	-70.5 ± 51.2	72.6

Table 4 reveals the distinct distribution patterns among emotion categories based on the distribution of MFCC coefficients (n=40) obtained from 900 emotion audio samples with balanced dataset (180 samples per class), where mean denotes the average value and CV is the coefficient of variation in percent. The most prominent pattern is the difference in mean values between emotion groups, where negative emotions (Disgust: -70.9, Disappointed: -70.5) have a much lower mean than other emotions (Neutral: -48.2, Happy: -48.4, Surprised: -47.8). In terms of variability, there are two interesting patterns: negative emotions tend to be more consistent (CV: Disgust 68.7%, Disappointed 72.6%) while positive and neutral emotions are more variable (CV: Neutral 94.0%, Happy 87.2%, Surprised 83.1%). The range of values also supports this finding where negative emotions show more extreme minimum values (up to -222.3) but with higher consistency. This finding suggests that negative emotions have more specific and consistent acoustic characteristics than positive and neutral emotions which show greater diversity of expression between individuals. This pattern indicates that MFCC features are able to distinguish well between emotions and thus provide an optimal contribution to BiLSTM-based emotion classification.

3.5 Model Training

Training was performed using the Adam optimizer with an initial learning rate of 0.001, batch size of 32, and a maximum of 150 epochs. An early stopping strategy with patience 20 was applied to prevent overfitting, while ReduceLROnPlateau decreased the learning rate when validation loss stagnated. The data was divided with a ratio of 80:20 using stratified split to maintain a balanced class distribution.

Table 5. BiLSTM Training Result

Epoch	Validation Accuracy	Validation Loss	Train Accuracy	Train Loss
1	30.6%	3.505	29.7%	3.799
25	67.2%	1.455	54.9%	1.721
50	76.7%	0.802	70.1%	0.941
100	91.7%	0.413	86.4%	0.484
132	93.3%	0.346	90.7%	0.347
150	91.7%	0.367	92.8%	0.275

Table 5 shows the training progress of the BiLSTM model which reached optimal convergence at the 132nd epoch with the highest validation accuracy of 93.3% and the lowest validation loss of 0.346. The comparison between training and validation metrics shows a healthy learning pattern, where the gap between training accuracy (90.7%) and validation accuracy (93.3%) at the best epoch is only 2.6%, indicating no significant overfitting. The consistent loss reduction from 3.505 at the first epoch to 0.346 at the 132nd epoch demonstrates the effectiveness of the applied optimization strategy, while the slight degradation at epoch 150 (validation accuracy drops to 91.7%) confirms the importance of implementing early stopping to prevent model performance degradation.

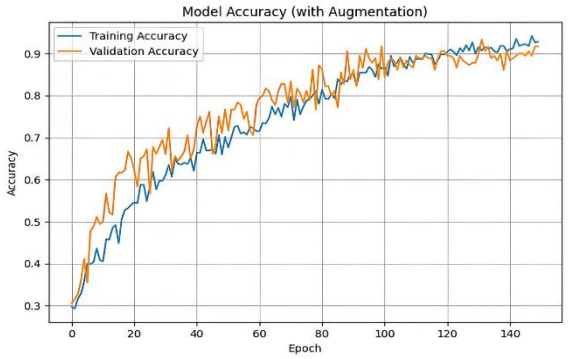

Figure 4. BiLSTM Training Accuracy Chart

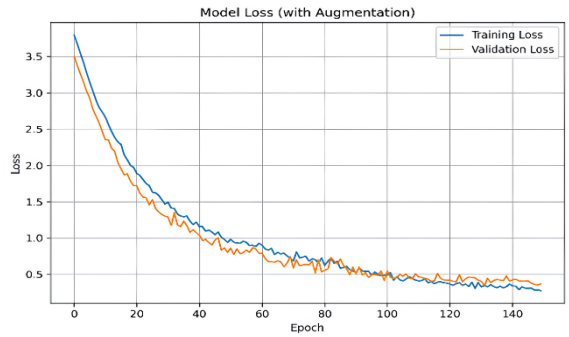

Figure 5. BiLSTM Training Loss Chart

Based on Figure 4 and Figure 5, it can be seen that the training accuracy and validation accuracy curves increase consistently from around 30% to over 90%, although there are occasional fluctuations in the validation accuracy which is above or below the training accuracy. However, the two curves still move in tandem without showing a significant widening distance, so there is no indication of serious overfitting. On the loss graph, both training loss and validation loss decrease steadily from an initial value of around 3.5 to reach 0.3-0.4 at the end of training, with the pattern converging to almost the same value. This shows that the learning process is stable and the implementation of data augmentation helps maintain a balance between training and validation performance.

Table 6. BiLSTM Classification Result

Class	Precision	Recall	F1-Score	Support
Disappointed	94.4%	94.4%	94.4%	36
Disgusted	89.2%	91.7%	90.4%	36
Happy	100%	88.9%	94.1%	36
Neutral	97.1%	91.7%	94.3%	36
Surprised	87.8%	100%	93.5%	36
Average	**93.7%**	**93.3%**	**93.3%**	**180**

Based on the evaluation results shown in Table 6, the emotion classification model performed very well with the highest precision in the Happy class (100%) and the lowest in the Surprised class (87.8%). For recall, the highest value was achieved in the Surprised class (100%) which means all the surprised samples were correctly detected, while the emotion "happy" had the lowest recall (88.9%) despite its perfect precision. Overall, the model achieved an average precision of 93.7%, recall of 93.3%, and F1-Score of 93.3% with a total accuracy of 93.3%, indicating that the model is effective in classifying different categories of emotions with consistent and balanced performance across all classes.

3.6 Confusion Matrix

The confusion matrix serves as an evaluation instrument that displays the distribution of model predictions against actual labels in matrix form. This visualization allows for an in-depth analysis of the model's ability to classify each emotion category, while also identifying patterns of error that occur during the prediction process.

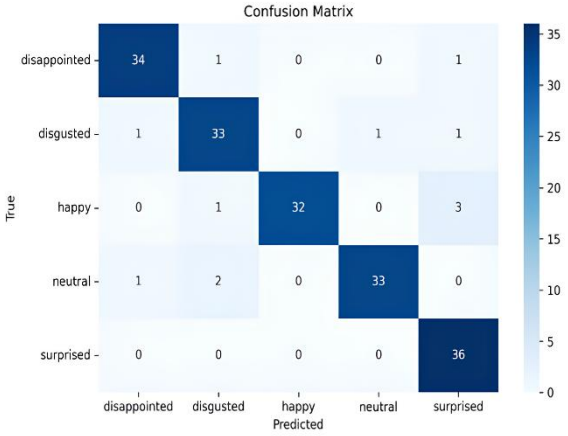

Figure 6. Confusion Matrix Result

The confusion matrix in Figure 6 demonstrates the impressive performance of the BiLSTM model with an overall accuracy of 93.33%. The balanced sample distribution with 36 samples each per class resulted in highly accurate predictions. The 'surprised' class achieved perfect performance with 100% accuracy (36/36 correct predictions), followed by 'disappointed' with 94.44% accuracy (34/36), 'disgusted' and 'neutral' each achieved 91.67% (33/36), while 'happy' obtained the lowest accuracy of 88.89% (32/36). The consistency of performance between classes shows that the model has successfully learned distinctive feature representations for each emotion category.

Although misclassification errors are minimal, with only 12 cases out of 180 samples, the error patterns still provide valuable insights. Each class experienced a maximum of 3 mispredictions, with an even distribution across categories. This low error rate indicates that the model has effectively extracted the acoustic features that distinguish each emotion. The minimal and evenly distributed error pattern indicates the robustness of the model in handling variability in voice emotion data.

Table 7. Comparison Of Model Performance

Researchers	Method	Object	Accuracy
Penelitian Bustamin et al.,2024 [6]	IndoWaveSentiment Dataset (Baseline)	Indonesian Emotion Audio	-
Penelitian Majiid et al., 2025 [7]	Random Forest & 45 Features	Indonesian Emotion Audio	90%
Penelitian Majiid et al., 2025 [7]	Gradient Boosting & 45 Features	Indonesian Emotion Audio	85%
Penelitian Majiid et al., 2025 [7]	Logistic Regression & 45 Features	Indonesian Emotion Audio	75%
Current Research	MFCC & BiLSTM	Indonesian Emotion Audio	93%

As shown in Table 7, the results demonstrate significant improvement over previous methods. The advantage of BiLSTM lies in its ability to capture bi-directional temporal dependencies in audio signals, providing a more comprehensive understanding of context than traditional methods such as Random Forest or Gradient Boosting.

```
Testing prediction function...
Prediction Results on Sample Files:
========================================
==========

File: 01-01-01-01.wav
True Emotion: neutral
Predicted: neutral | Confidence: 0.9974
Correct: ✓
All Probabilities:
  neutral: 0.9974
  disgusted: 0.0013
  disappointed: 0.0009
  surprised: 0.0003
  happy: 0.0001

File: 01-02-01-01.wav
True Emotion: happy
Predicted: happy | Confidence: 0.9817
Correct: ✓
All Probabilities:
  happy: 0.9817
  surprised: 0.0098
  disgusted: 0.0068
  neutral: 0.0014
  disappointed: 0.0004
```

Figure 7. Model Prediction Test Results

The prediction test results in Figure 7 demonstrate the BiLSTM model's ability to classify voice emotions with a very high confidence score. The model successfully predicted the emotion "Neutral" in the first sample with 99.74% confidence and the emotion "Happy" in the second sample with 98.17% confidence, both in accordance with the actual label. The probability distribution shows the dominance of

the predicted class with a very low probability of other classes, indicating the ability of the model to distinguish the MFCC feature characteristics distinctively. The high confidence score (>98%) in both samples reflects the effectiveness of the BiLSTM architecture in capturing the temporal patterns of Indonesian audio emotion signals.

4. CONCLUSION

This research successfully implemented an Indonesian speech emotion classification system using a combination of MFCC and BiLSTM with an accuracy of 93.33%. The audio augmentation technique proved effective in improving the robustness of the model through dataset diversification. The BiLSTM architecture shows superiority in capturing bidirectional temporal dependencies over conventional methods. The emotion "Surprised" shows the most distinctive characteristics with perfect accuracy. These results make a significant contribution to the development of speech-based emotion recognition systems for the Indonesian language and can be applied to various fields such as automated customer service, audio sentiment analysis, and speech-based interactive systems.

5. REFERENCES

[1] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, 2024.

[2] BPPTIK, "Voice Assistant AI: Pendamping Digital yang Siap Membantu," *bpptik.komdigi.go.id*, 2024. [Online]. Available: https://bpptik.komdigi.go.id/Publikasi/detail/voice-assistant-ai-pendamping-digital-yang-siap-membantu.

[3] MiiTel, "Survei: Indonesia Peringkat 4 Negara Paling Antusias dengan AI," *AI Analytics for Voice Communication*, 2024. [Online]. Available: https://miitel.com/id/survei-indonesia-peringkat-4-negara-paling-antusias-dengan-ai/.

[4] Y. K. Aini, T. B. Santoso, and T. Dutono, "Pemodelan CNN Untuk Deteksi Emosi Berbasis Speech Bahasa Indonesia," *J. Komput. Terap.*, vol. 7, no. 1, pp. 143–152, 2021.

[5] 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.

[6] A. Bustamin, A. M. Rizky, E. Warni, I. S. Areni, and Indrabayu, "IndoWaveSentiment: Indonesian audio dataset for emotion classification," *Data Br.*, vol. 57, 2024.

[7] M. R. N. Majiid, K. E. Setiawan, P. P. Yudha, A. Taufiq, and N. L. Setiawan, "Advancing Indonesian Audio Emotion Classification : A Comparative Study Using IndoWaveSentiment," vol. 7, no. 2, pp. 207–211, 2025.

[8] 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.

[9] Mariyanto and H. F. Pardede, "Exploring the Effectiveness of Deep Learning in Analyzing Review Sentiment," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 2, pp. 116–121, 2023.

[10] C. Zhang, H. Zhan, Z. Hao, and X. Gao, "Classification of Complicated Urban Forest Acoustic Scenes with Deep Learning Models," *Forests*, vol. 14, no. 2, 2023.

[11] F. F. Dias, M. A. Ponti, and R. Minghim, "Enhancing sound-based classification of birds and anurans with spectrogram representations and acoustic indices in neural network architectures," *Ecol. Inform.*, vol. 90, no. April, p. 103232, 2025.

[12] A. S. Kumar, T. Schlosser, S. Kahl, and D. Kowerko, "Improving learning-based birdsong classification by utilizing combined audio augmentation strategies," *Ecol. Inform.*, vol. 82, no. June, 2024.

[13] A. Alamsyah, F. Ardiansyah, and A. Kholiq, "Music Genre Classification Using Mel Frequency Cepstral Coefficients and Artificial Neural Networks: A Novel Approach," *Sci. J. Informatics*, vol. 11, no. 4, pp. 937–948, 2024.

[14] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," *Sci. Rep.*, vol. 15, no. 1, pp. 1–14, 2025.

[15] J. L. Bautista and Y. K. Lee, "Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation," pp. 1–14, 2022.

[16] E. Aurora Az Zahra, Y. Sibaroni, and S. Suryani Prasetyowati, "Classification of Multi-Label of Hate Speech on Twitter Indonesia using LSTM and BiLSTM Method," *JINAV J. Inf. Vis.*, vol. 4, no. 2, pp. 170–178, 2023.

[17] T. Li, "Optimizing the configuration of deep learning models for music genre classification," *Heliyon*, vol. 10, no. 2, p. e24892, 2024.

[18] F. Makhmudov, A. Kutlimuratov, and Y. I. Cho, "Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition," *Appl. Sci.*, vol. 14, no. 23, 2024.