Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 12, No. 3, September 2024, pp. 503~515
ISSN: 2089-3272, DOI: 10.52549/ijeei.v12i3.5662

Advanced Multimodal Emotion Recognition for Javanese Language Using Deep Learning

Fatchul Arifin[1], Aris Nasuha[1], Ardy Seto Priambodo[1], Anggun Winursito[1], Teddy Surya Gunawan[2]

[1]Department of Electronics and Informatics Engineering of Education, Universitas Negeri Yogyakarta, Indonesia
[2]Department of Electrical and Computer Engineering, International Islamic University Malaysia, Malaysia

Article Info

Article history:

Received Jun 23, 2024
Revised Aug 7, 2024
Accepted Aug 12, 2024

Keywords:

Javanese Emotion Recognition
Multimodal Deep Learning
Audio-Visual Integration
Emotion Detection Models
Cultural Emotion Analysis
Human-Computer Interaction

ABSTRACT

This research develops a robust emotion recognition system for the Javanese language using multimodal audio and video datasets, addressing the limited advancements in emotion recognition specific to this language. Three models were explored to enhance emotional feature extraction: the Spectrogram-Image Model (Model 1), which converts audio inputs into spectrogram images and integrates them with facial images for emotion labeling; the Convolutional-MFCC Model (Model 2), which leverages convolutional techniques for image processing and Mel-frequency cepstral coefficients for audio; and the Multimodal Feature-Extraction Model (Model 3), which independently processes video and audio features before integrating them for emotion recognition. Comparative analysis shows that the Multimodal Feature-Extraction Model achieves the highest accuracy of 93%, surpassing the Convolutional-MFCC Model at 85% and the Spectrogram-Image Model at 71%. These findings demonstrate that effective multimodal integration, mainly through separate feature extraction, significantly enhances emotion recognition accuracy. This research improves communication systems and offers deeper insights into Javanese emotional expressions, with potential applications in human-computer interaction, healthcare, and cultural studies. Additionally, it contributes to the advancement of sophisticated emotion recognition technologies.

Corresponding Author:

Fatchul Arifin,
Department of Electronic and Informatics Engineering,
Universitas Negeri Yogyakarta,
Yogyakarta, Indonesia.
Email: fatchul@uny.ac.id

1. INTRODUCTION

The examination of emotional expression through multimodal inputs is becoming increasingly important, particularly for the Javanese language, Indonesia's most widely spoken regional language. The complex and nuanced meanings of the Javanese emotional expression necessitate an approach that transcends the straightforward analysis of facial expressions or voice intonation [1]. The optimization of the extraction of human emotional characteristics and the enhancement of the accuracy of emotion recognition systems necessitate the integration of multimodal datasets, which combine a variety of modalities such as voice, text, and facial expressions [2-5].

In the rapidly evolving field of information technology, it is essential to have a more profound comprehension of human emotional expression, mainly to improve communication systems and advance artificial intelligence. With its profound cultural background and historical influence on daily life, Javanese plays a critical role in forming Indonesia's cultural identity [6, 7]. Its distinctive emotional discourse, distinct from other languages, reflects the enduring values, spiritual nuances, and local wisdom that have influenced the Javanese community for centuries. As a result, examining Javanese emotional expressions facilitated by

multimodal inputs is not merely an academic interest but a critical endeavor that can offer a more comprehensive understanding of the broader range of human emotion recognition technologies [8, 9].

The Javanese language, deeply embedded in the cultural fabric of Indonesia, carries a rich tapestry of emotional expressions that reflect the community's values, wisdom, and spiritual nuances. This language, spoken by millions, is characterized by its subtle and layered emotional discourse, where expressions of feelings such as "narima" (acceptance), "sabar" (patience), and "ngenes" (sadness) are not merely verbal but are intricately tied to facial expressions, tone of voice, and situational context. The complexity of Javanese emotional expressions, which often convey profound cultural meanings, necessitates an advanced and nuanced approach to emotion recognition that transcends simple auditory or visual analysis. Understanding these expressions is crucial for preserving the cultural heritage and linguistic uniqueness of the Javanese people and enhancing the development of more effective and culturally aware human-computer interaction systems [10, 11].

The capacity to detect and interpret human emotions has been significantly improved by recent developments in multimodal emotion recognition, which have integrated a variety of modalities, including speech, facial expressions, and text. For example, the RobinNet system, which employs an intermediate fusion of text and audio features, surpasses state-of-the-art solutions on benchmark datasets like IEMOCAP and MELD, recording a weighted accuracy of 72.8% [12]. Moreover, the development of additive angle penalty focus loss functions (APFL) has resolved concerns regarding fuzzy decision boundaries in speech emotion recognition, thereby illustrating the efficacy of multimodal and multitask learning frameworks [13]. These technological advancements emphasize the significance of integrating multiple sources of information to enhance the accuracy and robustness of emotion recognition systems across various languages and cultural contexts [14, 15].

The issue with recent emotion recognition systems is that they frequently depend on restricted features and modalities, insufficient to capture intricate and nuanced emotional expressions, particularly in culturally rich languages such as Javanese. This research resolves these constraints by implementing a multimodal approach to developing a dependable emotion recognition system for the Javanese language [16]. Among the research objectives are the development of a comprehensive multimodal dataset, designing and testing a variety of deep learning models, and assessing their ability to identify emotional expressions. This research aims to improve the system's effectiveness in real-world applications by promoting a comprehensive comprehension and interpretation of emotions by examining facial and vocal data [17].

The methodology involves creating a detection system using multimodal datasets to achieve these objectives. This process consists of the recording, preprocessing, and labeling audio and video data from Javanese speakers. The Spectrogram-Image Model converts audio inputs into spectrogram images combined with facial images for emotion labeling; the Convolutional-MFCC Model enhances feature extraction using convolution techniques for images and Mel-frequency cepstral coefficients (MFCC) for audio; and the Multimodal Feature-Extraction Model processes video and audio features separately before integrating their recognition results. The findings indicate that effective multimodal integration, mainly through separate feature extraction, significantly enhances emotion recognition accuracy. This research contributes to developing sophisticated emotion recognition technologies, with potential applications in human-computer interaction, healthcare, and cultural studies, providing deeper insights into Javanese emotional expressions and enhancing communication systems' effectiveness.

2. RESEARCH METHODOLOGY

Beginning with an exploratory phase, the research comprehensively evaluates the current state of emotion recognition systems in the field. This initial investigation aims to clarify the current system architectures and their functionalities, thereby identifying potential areas for improvement. The ultimate objective is an advanced system that substantially enhances existing standards. This systematic approach guarantees a thorough comprehension of the fundamental technologies and frameworks essential for developing innovative solutions to advance the development of more advanced emotion recognition systems.

The Javanese language's emotion detection domain is particularly dependent on data, a critical component of artificial intelligence research. Currently, this field lacks standardized datasets specifically designed for emotion recognition. Consequently, developing a comprehensive database—which encompasses the recording, preprocessing, and labeling of data—is necessary for this research. This database will facilitate exploring and implementing the most appropriate multimodal emotion recognition models. To enhance the precision and effectiveness of emotion detection technologies, it is imperative to compile and organize this data systematically.

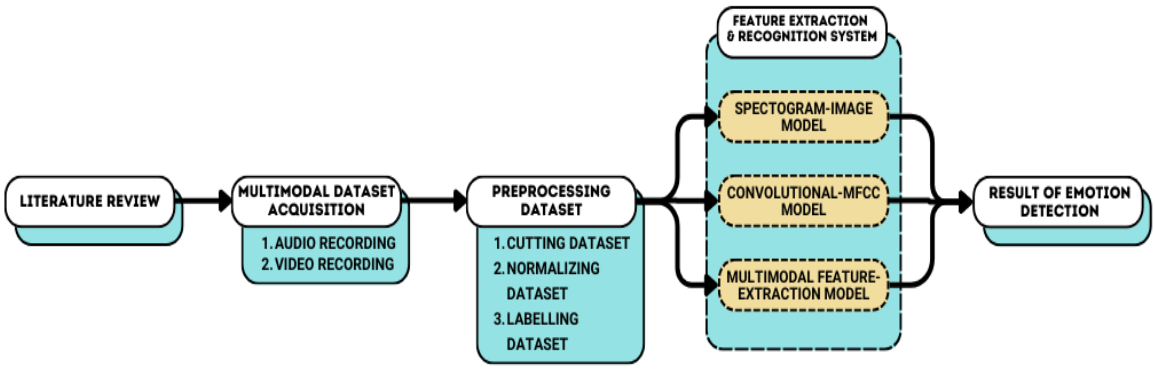


Figure 1. Research Methodology

The development of a dataset is a critical initial step that follows an exhaustive review of pertinent literature. The designed system is trained and validated by collecting multimodal input data, including audio and video. Following the design of the emotion recognition system, the process progresses to its implementation and concludes with system testing. The system's reliability and efficacy in real-world applications are contingent upon completing these stages. Fig. 1 illustrates the sequential stages of the research, which illustrates the structured approach to developing a robust emotion recognition system that integrates multiple data sources for improved performance.

2.1. Multimodal Dataset Acquisition

The dataset compiled for this emotion recognition study specifically intends to include detailed video and audio data annotated with emotional information to facilitate the training and testing phases of the proposed models. The methodology used in data collection is depicted in Fig. 2, which outlines the steps involved in acquiring and preparing the data for subsequent algorithmic processing.

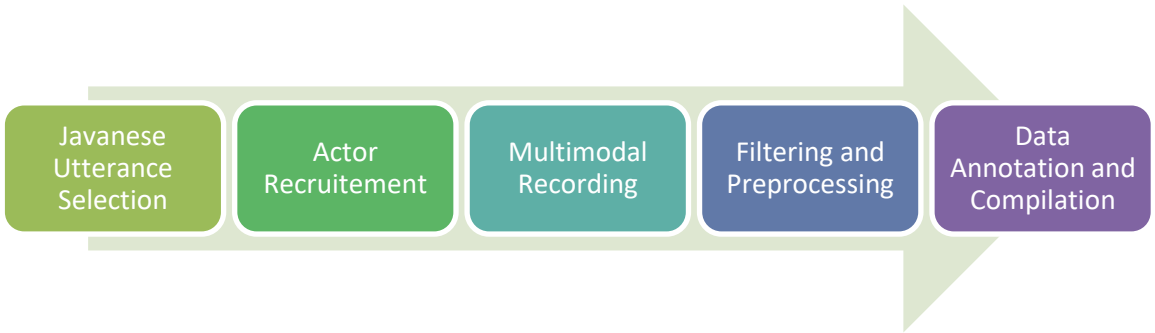


Figure 2. Multimodal Data Collection Processes

The three most widely recognized methods for constructing an emotion recognition database are natural, elicited, and actor-simulated recordings. A high probability of success was the primary reason for selecting the actor-simulated method in this investigation [10, 16]. To improve the reliability and emotional range of the data, Kamasetra UNY, an extracurricular arts and theater organization at Universitas Negeri Yogyakarta (UNY), recruited experienced actors. The cast consisted of ten Javanese speakers, with a gender balance of five males and five females. Each actor pronounced phrases corresponding to specific emotions seven times, resulting in a substantial dataset of 1,680 voice recordings and an equal number of video recordings.

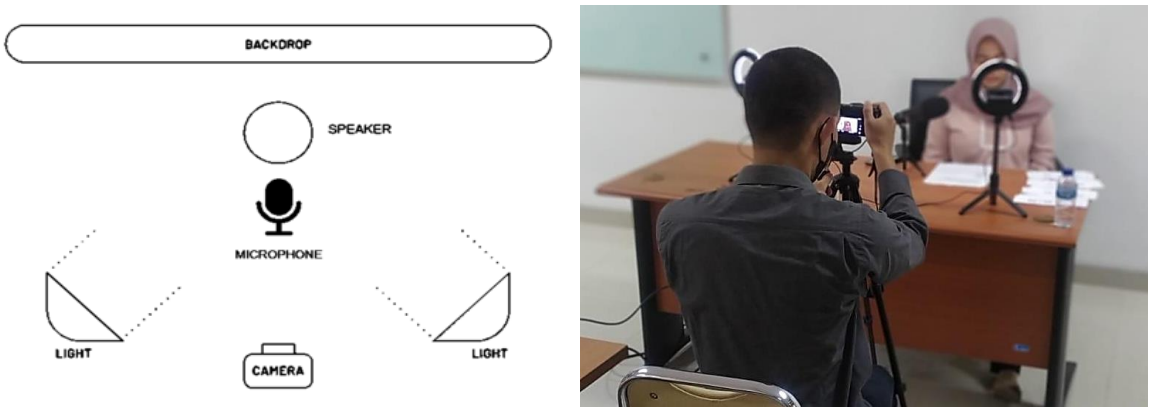
The actors exhibited six distinct emotions during the recordings: happiness, sadness, fear, anger, neutrality, and surprise. Javanese language experts assisted in selecting the sentences articulated in Javanese from classical Javanese literary works, such as novels, books, and dramas. The sentences selected were as follows:

- "*Yen wis entuk jeneng, bakal entuk jenang*" (If you have earned trust, you will benefit)
- "*Kabeh kuwi wis diatur sing gawe urip*" (Everything is arranged by God)
- "*Wis ngomong wae rasah sungkan*" (Just say it, no need to be ashamed)
- "*Aku pengine diakui merga karyaku*" (I want to be recognized for my achievements)

The variables chosen for the multimodal dataset in this study were carefully selected to balance data diversity and practical feasibility. Including 10 participants was sufficient to provide a varied yet manageable dataset, capturing a wide range of emotional expressions without overwhelming the data collection process. Having each participant perform seven repetitions of each sentence ensures a robust sample size for each emotion and sentence combination, aiding in the reliability and statistical validity of the model training and testing phases.

The selection of six distinct emotions—happiness, anger, neutral, sadness, fear, and surprise—covers a comprehensive spectrum of primary emotions commonly studied in emotion recognition research, providing a solid foundation for analyzing emotional expressions. Meanwhile, choosing four sentences allows for a varied linguistic context, which helps the models learn to generalize emotional expressions across different verbal content. This methodological approach, incorporating a balanced number of participants, repetitions, emotions, and sentences, ensures the creation of a rich and diverse dataset suitable for developing and evaluating sophisticated emotion recognition systems.

This methodological approach guarantees that the recordings accurately represent a wide variety of emotional expressions pertinent to the cultural and linguistic contexts, thereby enabling the detailed analysis and identification of emotions in the spoken language of Javanese. As illustrated in Fig. 3, audio and video recordings were conducted in a standardized environment free of noise, echoes, and other auditory disturbances. The configuration comprised a single camera and two lighting units to guarantee optimal visibility. Clear audio was captured using an advanced unidirectional condenser microphone connected to a computer with noise-cancellation software (Krisp AI). The audio data was recorded in mono-channel format using Audacity software, which was configured to a sampling rate of 16 kHz. This ensured that the recordings were high quality, essential for accurately analyzing emotion recognition research.



(a) Multimodal Recording System Setup (b) Multimodal Recording Process at Laboratory

Figure 3. Multimodal Recording Processes

Table 1. Multimodal Dataset Description

Item	Description
Number of Emotions	6 (anger, fear, happiness, sadness, neutral, surprise)
Number of Speakers	10 (five men and five women)
Repetition	7 times
Video Format	RAW, RGB color mode, 1440×1080 image resolution, 30 fps
Audio Format	WAV, 16 kHz, 512 kbps
Data per Emotion	10 (actors) × 4 (sentences) × 7 (repetitions) = 280 per emotion
Total Data	1680 voice recordings and 1680 video recordings

After data acquisition, the dataset was prepared for analysis by implementing several preprocessing procedures. This procedure involved normalizing the recordings' length, applying appropriate labeling, and trimming the data to relevant segments. Furthermore, the recorded audio files were converted to the WAV format at 512 kbps to standardize the audio quality and enable uniform processing. Table 1 summarizes the multimodal dataset's specifications in detail. The dataset guarantees a solid foundation for developing and testing multimodal emotion recognition systems designed to accommodate Javanese emotional expressions' subtleties by adhering to these structured methodologies.

2.2. Data Labeling

Audio and video files were immediately assigned specific names during the recording process using a codification system to ensure precise labeling. This system followed the format $\mathbf{X_1X_2X_3X_4}$, where each variable represented a different aspect of the recording, as follows:

- $\mathbf{X_1}$ identified the artist (with 10 artists numbered from 1 to 10).
- $\mathbf{X_2}$ denoted the type of emotion being expressed (with six possible emotions: 0 for Happiness, 1 for Anger, 2 for Neutral, 3 for Sadness, 4 for Fear, and 5 for Surprise).
- $\mathbf{X_3}$ indicated the sentence being read (with four different sentences available).
- $\mathbf{X_4}$ specified the repetition count (each sentence was repeated seven times).

For example, code 2122 signified that the second artist was expressing anger, reading the second sentence, and it was the second repetition. This rigorous labeling process systematically organized the data, making it easily identifiable and accessible for further analysis. To ensure the accuracy of the emotional expressions, a Kamasetra supervisor was present during the recording sessions to validate the artists' actions, confirming that the emotions conveyed matched the intended expressions. This immediate and supervised labeling process provided high accuracy and validity in the emotional data, establishing a reliable foundation for subsequent analysis. Fig. 4 presents a sample of the image data, showcasing the diverse emotional expressions captured during the study.

Figure 4. Image Samples of Various Emotions

2.3. Proposed Multimodal Recognition Systems

This research aims to develop a multimodal dataset that incorporates audio and video inputs to custom-design an emotion recognition system for the Javanese language. The Spectrogram-Image Model, the Convolutional-MFCC Model, and the Multimodal Feature-Extraction Model were each developed to accomplish this. To ascertain the most effective strategy for emotion recognition in a culturally specific context, each model was selected to investigate distinct methods of processing and integrating multimodal data.

The necessity of comparing and contrasting various methods of multimodal emotion recognition for the Javanese language is the basis for selecting these three experimental scenarios. The Spectrogram-Image Model employs an innovative method of combining facial images from video inputs with audio signals and converting them into spectrogram images. This model utilizes CNNs to simultaneously analyze patterns and textures in sound and visual data, thereby enabling the comprehensive extraction of features and the recognition of emotions in a unified process.

In contrast, the Convolutional-MFCC Model is intended to improve feature extraction by utilizing convolutional techniques on video data and Mel-Frequency Cepstral Coefficients (MFCC) on audio data. This dual-modality approach guarantees precise emotion recognition by capturing essential auditory characteristics and high-level visual features. This model's comprehensive analysis enables a more profound comprehension of the visual and auditory cues essential for identifying emotions in the Javanese language.

The Multimodal Feature-Extraction Model employs an alternative approach by independently processing audio and video inputs before combining their recognition outcomes. This model enhances recognition accuracy by optimizing the strengths of each modality, enabling specialized, independent processing. By combining detailed visual and auditory emotional cues, this model ensures that both modalities fully contribute to the recognition process, thereby addressing the complexities of Javanese emotional expressions.

It is imperative to conduct a comparative analysis of these three models to determine the most effective approach to multimodal emotion recognition. This research not only helps to advance the development of precise emotion recognition systems but also aids in creating culturally sensitive technologies that comprehend and respect the subtleties of the Javanese language.

2.3.1. Spectrogram-Image Model (Model 1)

The Spectrogram-Image Model transforms audio inputs into spectrogram images, merging them with facial images from video inputs. As shown in Fig. 5, each video segment is segmented into distinct images, and audio data is converted into histogram images. These images undergo size normalization before being combined into a composite image annotated with the associated emotional content. Feature extraction and emotion recognition are then performed using a CNN. This model integrates audio and visual data into a single image-based format, enabling the simultaneous processing of multimodal information. This innovative approach simplifies the data fusion process, ensuring robust emotion recognition through a unified representation of audio-visual inputs.

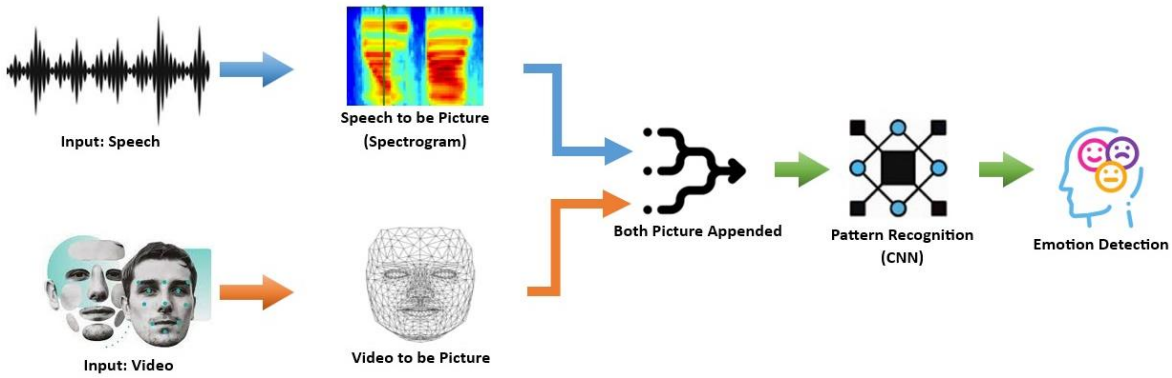


Figure 5. Proposed Spectrogram-Image Model (Model 1)

The Spectrogram-Image Model is compelling because of its innovative approach to combining audio and video data. Instead of converting audio signals into spectrogram images, the model seamlessly integrates them with video-derived images. This approach leverages the visual representation of audio frequencies to effectively exploit spatial and spectral features in CNN. By converting audio inputs into spectrogram images, the model leverages the visual processing capabilities of CNNs to identify patterns and textures in sound, thereby improving the system's capacity to identify emotional nuance in speech. Concurrently, video frames are processed to capture critical visual cues from facial expressions. This composite input stream, which combines spectrogram and image data, enables the model to perform comprehensive feature extraction and recognition, thereby significantly increasing its ability to detect emotions. This method bridges the gap between auditory and visual data. It improves the model's robustness in managing multimodal inputs, thereby establishing a superior framework for accurate and reliable recognition of emotions in the Javanese language.

2.3.2. Convolutional-MFCC Model (Model 2)

The Convolutional-MFCC Model enhances feature extraction using advanced techniques for both images and audio. As shown in Fig. 6, video inputs are segmented into multiple images and processed using convolutional algorithms to extract relevant features. Simultaneously, audio inputs are analyzed using the Log-Mel Filterbank Extraction Method (MELL), a robust technique for capturing essential audio characteristics. The extracted features from video and audio sources are integrated into a composite dataset, annotated with labels corresponding to the depicted emotions. This model leverages the strengths of convolutional processing for visual data and MFCC for audio data, ensuring precise emotion recognition through a comprehensive analysis of multimodal inputs.

Leveraging the strengths of CNNs and MFCC, the Convolutional-MFCC Model provides a robust approach to multimodal emotion recognition by integrating advanced feature extraction techniques for audio

and video inputs. This model utilizes CNNs to extract high-level features from video data, capturing subtle visual cues and intricate facial expressions. Concurrently, it employs MFCC to extract critical auditory characteristics from speech, thereby accurately representing the power spectrum of sound essential for identifying emotional tones. The model generates a comprehensive dataset that offers a multifaceted, enriched comprehension of emotional expressions by incorporating these detailed features from both modalities. By combining these features, CNN can perform sophisticated pattern recognition, improving the model's accuracy and reliability in emotion detection. This architecture is particularly effective in identifying the nuanced and culturally specific emotional expressions in the Javanese language, as it not only optimizes the information extracted from each modality but also guarantees that the combined data provides a more comprehensive understanding of the emotional context.

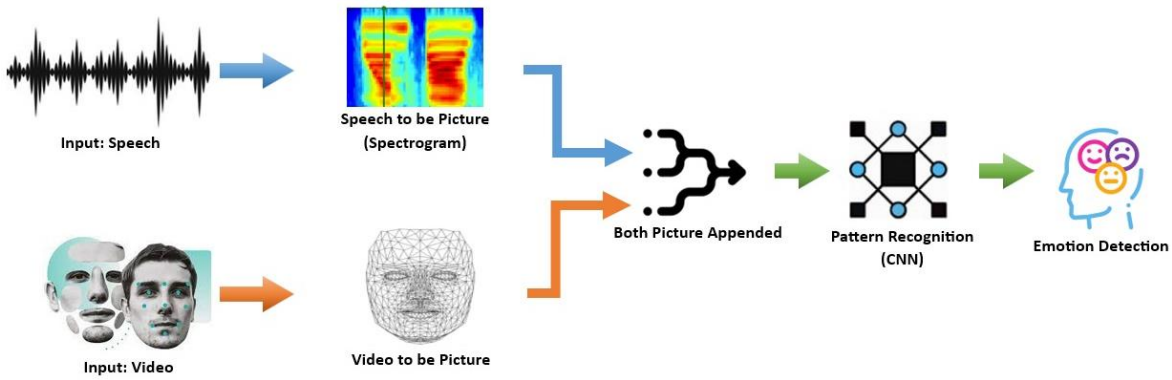


Figure 6. Proposed Convolutional-MFCC Model (Model 2)

2.3.3. Multimodal Feature-Extraction Model (Model 3)

The Multimodal Feature-Extraction Model processes video and audio inputs separately before integrating their recognition results. As shown in Fig. 7, this model first extracts features from video segments, converting each video into multiple image frames. These frames undergo feature extraction using CNNs. Concurrently, audio features are extracted using MFCC and analyzed with an Artificial Neural Network (ANN). The final emotion classification is determined by integrating the results from both modalities, prioritizing the higher value between video and audio recognitions. This approach ensures that visual and auditory emotional cues are effectively captured and analyzed independently before combining their insights.

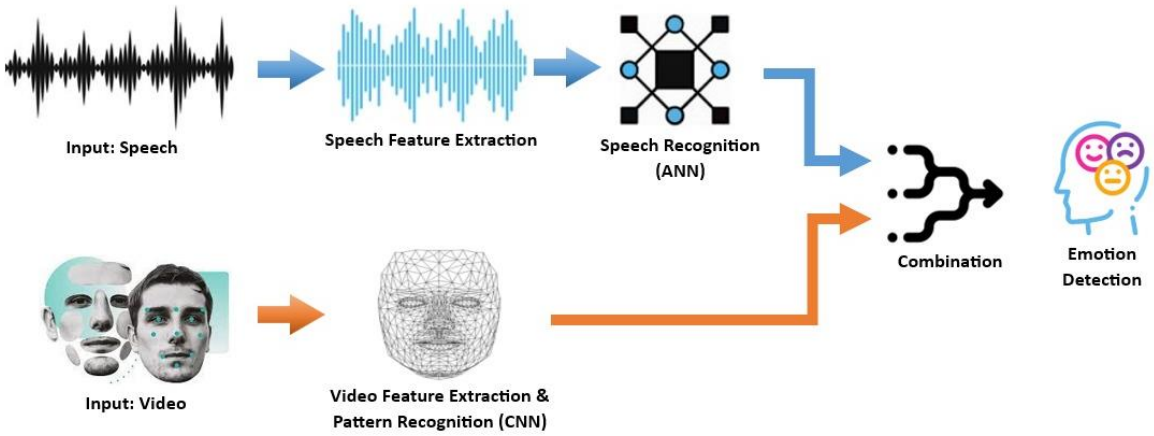


Figure 7. Proposed Multimodal Feature-Extraction Model (Model 3)

By utilizing specialized processing for each modality, the Multimodal Feature-Extraction Model's deep learning architecture guarantees the extraction of high-fidelity and robust features from audio and video inputs. The model effectively captures intricate spatial patterns in facial expressions by CNNs for video data. MFCC and ANNs ensure precise representation and analysis of speech intonation and emotional tone for audio data. This independent yet complementary processing optimizes the strengths of each modality, resulting in a more comprehensive and precise emotion recognition system. The subsequent integration of these features enables a comprehensive comprehension of the emotional context, thereby significantly improving recognition accuracy by combining the detailed visual and auditory emotional cues. A robust solution for emotion

detection, this dual-modality approach not only addresses the limitations of unimodal systems but also ensures that the unique and culturally rich emotional expressions inherent in the Javanese language are effectively captured and interpreted.

The multimodal emotion recognition methodologies demonstrate strategic improvements as they progress from the Spectrogram-Image Model (Model 1) to the Convolutional-MFCC Model (Model 2) and ultimately to the Multimodal Feature-Extraction Model (Model 3). Model 1 establishes a foundational approach by transforming audio inputs into spectrogram images and combining them with visual data, leveraging the capabilities of CNNs to process both types of data within a unified framework simultaneously. This method capitalizes on the visual representation of audio frequencies, enhancing the model's ability to identify patterns across modalities. Model 2 advances this concept by utilizing MFCC for detailed audio feature extraction and applying CNNs for sophisticated video analysis. By integrating these high-level features into a comprehensive dataset, Model 2 improves pattern recognition capabilities. Finally, Model 3 ensures the most robust performance by independently processing audio and video inputs before integrating the recognition results. This independent processing allows for specialized extraction and recognition of features from each modality, minimizing interference and maximizing the potential of each data type. The progression from Model 1 to Model 3 illustrates a thoughtful refinement in the architecture, with each step increasing the complexity and effectiveness of feature extraction and integration processes. This systematic enhancement results in a more accurate and comprehensive emotion recognition system that is adept at capturing the nuanced emotional expressions of the Javanese language.

3. RESULTS AND DISCUSSION

Model 1 (Spectrogram-Image Model), Model 2 (Convolutional-MFCC Model), and Model 3 (Multimodal Feature-Extraction Model) were each developed to capitalize on unique aspects of multimodal data processing and integration. To process both data types concurrently, the spectrogram image model converts audio inputs into spectrogram images and combines them with visual data, utilizing CNNs. The Convolutional-MFCC model optimizes feature extraction by integrating high-level features into a comprehensive dataset by employing convolution techniques on images and MFCC for audio. The Multimodal Feature-Extraction Model processes video and audio features separately before integrating their recognition results, enabling specialized extraction and recognition from each modality.

3.1. Experimental Setup

Robust hardware and sophisticated software are essential components of the experimental setup that manages the computational demands and intricate data processing tasks associated with developing and testing the multimodal emotion recognition system. Table 2 provides a comprehensive overview of the hardware and software specifications. We employed an NVidia Tesla T4 16 GB GPU to accelerate deep-learning computations. Python 3 is the software environment's foundation, incorporating indispensable libraries, including OpenCV for image processing, librosa for audio analysis, and TensorFlow for developing Deep Learning models. This comprehensive setup guarantees efficient data handling, high-performance processing, and accurate model training, establishing a strong foundation for the research's sophisticated analytical needs.

Table 2. Hardware and Software Used

Item	Description
Hardware	Intel Core i5 12400F, DDR4 16 GB, SSD NVme 1 TB, NVidia Tesla T4 16 GB
Software	Python 3, OpenCV, os, librosa, numpy, math, Scikit-Learn, MoviePy, Keras, joblib, glob, soundfile, pandas, TensorFlow, matplotlib, seaborn, skimage

Table 3, which provides a comprehensive overview of the model architectures and trainable parameters, is essential for comprehending the computational complexity and design of each model employed in this investigation. The Convolutional-MFCC, Spectrogram-Image, and Multimodal Feature-Extraction Models all employ CNNs to process multimodal data. However, their architectures are specifically designed to optimize distinct aspects of emotion recognition. We utilized an auto-tuning mechanism to optimize the model performance and adjust critical hyperparameters such as the learning rate, epoch, and batch size. This approach was chosen due to the impracticality of manually testing the many possible combinations, which could extend into the thousands. Auto-tuning allowed us to efficiently identify the optimal values for these parameters, ensuring the best performance of the models without the exhaustive trial-and-error process that manual tuning would require.

The Spectrogram-Image Model utilizes convolutional layers to convert audio inputs into spectrogram images combined with visual data. This method capitalizes on the spatial processing capabilities of CNNs, allowing the model to accurately identify patterns in visual and audio inputs. The architecture consists of three

Conv2D layers with progressively larger filter sizes, followed by MaxPooling layers to mitigate overfitting and reduce spatial dimensions. Reflecting the complexity necessary to manage the transformation and integration of multimodal data, this model has 8,482,304 trainable parameters.

The Convolutional-MFCC Model improves feature extraction by employing convolution techniques for video data and MFCC for audio data. This model utilizes a CNN architecture comparable to the Spectrogram-Image Model; however, it processes three-channel (RGB) images, slightly increasing the number of trainable parameters to 8,482,880. With the incorporation of MFCC for audio, the model can more effectively capture essential audio features, increasing emotion recognition accuracy. The heightened parameter count is indicative of the heightened complexity that is associated with the integration of intricate audio and video features.

The Multimodal Feature-Extraction Model processes audio and video features independently before integrating their recognition results. This method optimizes the potential of each modality by enabling specialized extraction and recognition before combining the results. The architecture of the CNN for video processing is identical to that of the Convolutional-MFCC Model. It consists of three Conv2D layers, followed by MaxPooling, and a final Dense layer intended for classification. Around 8,482,880 parameters are trainable in total. The architecture of this model underscores the necessity of distinct processing for multimodal inputs to optimize the extraction of pertinent features from each modality and reduce interference.

Table 3. Model Architectures and Trainable Parameters

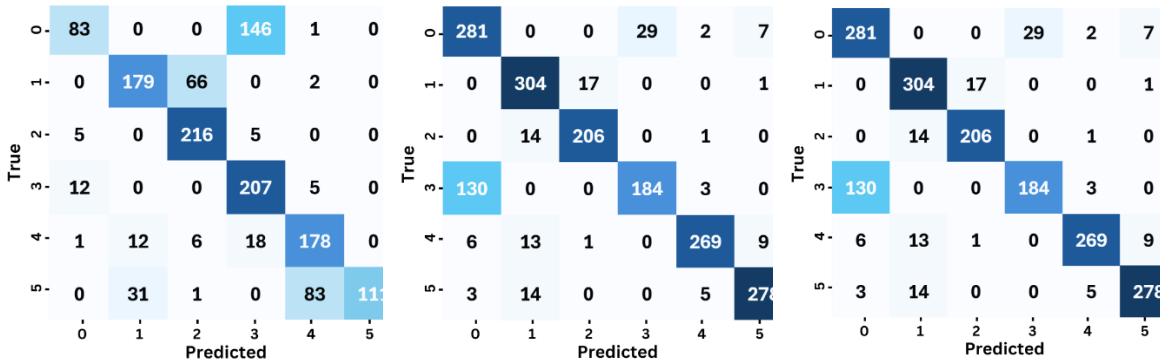
Model	Architecture Details	Number of Trainable Parameters
Spectrogram-Image (Model 1)	Conv2D(32, (3, 3), input_shape=(height, width, 1)) → MaxPooling2D((2, 2)) Conv2D(64, (3, 3)) → MaxPooling2D((2, 2)) Conv2D(128, (3, 3)) → MaxPooling2D((2, 2)) Flatten() → Dense(1024)	8,484,304
Convolutional-MFCC (Model 2)	Conv2D(32, (3, 3), input_shape=(height, width, 3)) → MaxPooling2D((2, 2)) Conv2D(64, (3, 3)) → MaxPooling2D((2, 2)) Conv2D(128, (3, 3)) → MaxPooling2D((2, 2)) Flatten() -> Dense(1024)	8,482,880
Multimodal Feature-Extraction (Model 3)	Conv2D(32, (3, 3), input_shape=(height, width, 3)) -> MaxPooling2D((2, 2)) Conv2D(64, (3, 3)) → MaxPooling2D((2, 2)) Conv2D(128, (3, 3)) → MaxPooling2D((2, 2)) Flatten() -> Dense(1024)	8,482,880

The dataset utilized in this study was obtained through our recordings, featuring 10 participants comprising 5 males and 5 females. The recordings captured six distinct emotions, and each participant articulated four different sentences, each repeated seven times. This process resulted in a comprehensive dataset totaling 1,680 recordings. Of this data, 70% was allocated for training purposes, while the remaining 30% was reserved for testing.

3.2. Performance Evaluation of Various Models

The performance of the three deep learning models—Spectrogram-Image Model (Model 1), Convolutional-MFCC Model (Model 2), and Multimodal Feature-Extraction Model (Model 3)—was evaluated based on their ability to recognize emotions from multimodal inputs accurately. The confusion matrices presented in Fig. 8, along with the precision, recall, F1-scores in Table 4, and overall accuracy in Table 5, provide a comprehensive overview of each model's effectiveness.

Model 1 demonstrated an overall accuracy of 71.15%, as shown in Table 5. The confusion matrix in Figure 7(a) reveals that Model 1 struggled with certain emotions, particularly "happiness" (label 0) and "fear" (label 4), where significant misclassifications were observed. The precision and recall values for "happiness" were 0.82 and 0.36, respectively, indicating that while the model could identify instances of happiness, it frequently misclassified other emotions as happiness. Similarly, the model's performance on "fear" was suboptimal, with a precision of 0.63 and a recall of 0.83, suggesting a high rate of false positives. These shortcomings highlight the limitations of Model 1 in effectively integrating audio and visual data early in the processing pipeline, which may lead to the loss of crucial information.



(a) Spectrogram-Image (b) Convolutional-MFCC Multimodal Feature-Extraction

Figure 8. Confusion Matrix for Various Models

Table 4. Performance of Various Deep Learning Architectures

Emotion	Precision			Recall			F1-score		
	M1	M2	M3	M1	M2	M3	M1	M2	M3
0 (happiness)	0.82	0.67	0.82	0.36	0.88	0.97	0.50	0.76	0.89
1 (anger)	0.77	0.89	0.97	0.72	0.94	0.96	0.75	0.92	0.96
2 (neutral)	0.73	0.91	0.98	0.96	0.93	0.97	0.83	0.92	0.98
3 (sadness)	0.53	0.86	0.93	0.92	0.58	0.88	0.67	0.70	0.91
4 (fear)	0.63	0.97	0.95	0.83	0.90	1.00	0.72	0.93	0.98
5 (surprise)	1.00	0.93	0.98	0.49	0.93	0.83	0.66	0.93	0.90
Weighted Average	0.75	0.85	**0.94**	0.71	0.85	**0.93**	0.69	0.85	**0.93**

* M1: Spectrogram-Image Model, M2: Convolutional-MFCC Model, M3: Multimodal Feature-Extraction Model

Table 5. Accuracy of Various Deep Learning Architectures

Model	Accuracy
Spectrogram-Image	0.7115
Convolutional-MFCC	0.8546
Multimodal Feature-Extraction	**0.9332**

Model 2 exhibited a significantly higher overall accuracy of 85.46%, as noted in Table 5. The confusion matrix in Figure 7(b) shows improved performance across most emotion categories compared to Model 1. This model particularly excelled in recognizing "anger" (label 1) and "neutral" (label 2), with precision values of 0.89 and 0.91 and recall values of 0.94 and 0.93, respectively. The higher precision and recall scores indicate that Model 2 was more effective in accurately classifying these emotions, likely due to the enhanced feature extraction capabilities provided by the MFCC for audio inputs and CNN for video inputs. The model's robustness in handling multimodal data can be attributed to its effective combination of detailed audio and visual features before emotion recognition.

Model 3 outperformed both Model 1 and Model 2, achieving an impressive overall accuracy of 93.32%, as detailed in Table 5. The confusion matrix in Figure 7(c) highlights the model's superior performance, particularly in recognizing "happiness" (label 0) and "fear" (label 4), with precision values of 0.82 and 0.95 and recall values of 0.97 and 1.00, respectively. The high F1 scores across all emotion categories indicate that Model 3 effectively minimizes false positives and false negatives, providing a balanced and accurate emotion recognition system. The independent processing of audio and video inputs, followed by the integration of recognition results, allows Model 3 to leverage the strengths of each modality without interference, leading to more precise emotion classification.

3.3. Discussion

The comparative analysis of the three models—Spectrogram-Image, Convolutional-MFCC, and Multimodal Feature-Extraction—demonstrates a clear strategic enhancement in handling multimodal data for emotion recognition, particularly within the Javanese language context. While innovative, Model 1's early fusion approach faced significant challenges in preserving essential information from each modality, leading to lower accuracy and higher misclassification rates. In contrast, Model 2 showed a marked improvement in accuracy by effectively combining detailed audio and visual features before recognition. However, it still encountered difficulties achieving optimal precision and recall across all emotions.

Model 3's superior performance underscores the critical importance of separate modality processing. Model 3 minimizes interference between modalities and maximizes the extraction of relevant features by allowing each modality to be processed independently before integrating the results. This approach enhances

the model's accuracy and ensures a more robust and reliable emotion recognition system, mainly when dealing with the nuanced emotional expressions inherent in the Javanese language. The high precision, recall, and F1 scores achieved by Model 3 across all emotion categories validate its effectiveness and justify its adoption for advanced emotion recognition tasks.

The results of this study represent a significant advancement in multimodal emotion recognition, particularly with the Multimodal Feature-Extraction Model achieving an accuracy of 93.32%. This model's independent processing of audio and video inputs before integration is a key factor contributing to its high performance. When compared to prior research, our results stand out. For example, a 72.8% accuracy using the MELD dataset with an intermediate fusion of text and audio features was achieved in [2]. Similarly, the effectiveness of various fusion strategies was emphasized in [18], such as early, late, and hybrid fusion, in improving multimodal emotion recognition accuracy. These studies highlight the critical role of multimodal integration in enhancing the robustness of emotion recognition systems, and our findings further reinforce this approach, especially in the context of the culturally nuanced Javanese language.

In addition, our research aligns with findings from comparative studies [2, 19], which emphasizes the significance of advanced AI technologies in improving emotion recognition accuracy. These studies support our approach of effective multimodal integration to interpret complex emotional states from audiovisual inputs. The use of CNNs and MFCC, as implemented in Model 2, reflects recent advancements in processing and analyzing complex data structures [20]. This is crucial in handling the intricacies of Javanese emotional expression, where cultural subtleties play a significant role, paralleling findings from research on other regional languages. Moreover, the unique cultural context of Javanese emotion, with expressions like "narima" (acceptance) and "ngenes" (sadness), underscores the need for culturally sensitive AI models. Studies considering cultural factors in AI development [11] resonate with our findings, suggesting that understanding these nuances is imperative for creating humane and intuitive AI technologies. Our research supports existing theories on multimodal emotion recognition systems and provides new insights into their application within culturally diverse settings. The high performance of our models, particularly Model 3, highlights the potential for future research and development in emotion recognition technologies, emphasizing the need for robust, context-aware systems capable of interpreting the subtle complexities of human emotions.

Our research supports existing theories on multimodal emotion recognition systems and provides new insights into their application within culturally diverse settings. The high performance of our models, particularly Model 3, highlights the potential for future research and development in emotion recognition technologies, emphasizing the need for robust, context-aware systems capable of interpreting the subtle complexities of human emotions.

4. CONCLUSION

By developing and evaluating three distinct deep learning models—the Spectrogram-Image Model, Convolutional-MFCC Model, and Feature-Extraction Model—this research presents significant advancements in multimodal emotion recognition for the Javanese language. The results underscore the effectiveness of processing audio and video inputs separately before integrating their recognition results, as the Multimodal Feature-Extraction Model achieved the highest accuracy of 93.32%. This highlights the superior performance of the model. This method is essential for accurately interpreting the nuanced emotional expressions inherent in the Javanese language, as it minimizes interference and maximizes feature extraction. Comparative analyses with existing research have confirmed the significance of culturally sensitive models and sophisticated AI technologies in improving emotion recognition systems. Future research will expand the dataset to encompass a more diverse array of emotional expressions and speakers and investigate real-time emotion recognition applications. Furthermore, integrating contextual information and enhancing model interpretability will be prioritized further to refine the accuracy and reliability of emotion recognition systems, thereby facilitating more intuitive and human-like AI interactions. These endeavors aim to further emotion recognition and its applications in various fields, such as healthcare, human-computer interaction, and cultural studies.

ACKNOWLEDGMENTS

The authors wish to express their gratitude and appreciation to all those who contributed to the successful development of the Javanese Multimodal Emotion Recognition System. First, we sincerely thank Universitas Negeri Yogyakarta and International Islamic University Malaysia for providing the financial support necessary for this research. Second, we are grateful to Kamasetra UNY for their willingness to participate as actors in our data collection process.

REFERENCES

[1] N. Ahmed, Z. Al Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intelligent Systems with Applications*, vol. 17, p. 200171, 2023.

[2] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.
[3] A. Ashraf, T. S. Gunawan, F. Arifin, M. Kartiwi, A. Sophian, and M. H. Habaebi, "On the Audio-Visual Emotion Recognition using Convolutional Neural Networks and Extreme Learning Machine," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 10, no. 3, pp. 684-697, 2022.
[4] A. Ashraf, T. S. Gunawan, F. Arifin, M. Kartiwi, A. Sophian, and M. H. Habaebi, "Enhanced Emotion Recognition in Videos: A Convolutional Neural Network Strategy for Human Facial Expression Detection and Classification," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 11, no. 1, pp. 286-299, 2023.
[5] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795-47814, 2021.
[6] A. S. Jamiluddin, S. K. Udja, and R. Safithri, "Meaning and Message of Communication Behaviour of Javanese Ethnic Traders to Prospective Buyers," in *International Conference on Halal, Policy, Culture and Sustainability Issues*, 2022, vol. 4, no. 1, p. 19.
[7] P. Wijonarko and A. Zahra, "Spoken language identification on 4 Indonesian local languages using deep learning," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 6, pp. 3288-3293, 2022.
[8] E. T. Sulistyo, "Emotional Intelligence And Balanced Personality In Javanese Cultural Understanding," *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 18, no. 4, pp. 3344-3359, 2021.
[9] S. A. Kumala, "Analysis of Language Attitude and Language Preservation in Javanese Language.: A Case Study of Javanese Speaker in Madiun, East Java," *e-LinguaTera*, vol. 1, no. 1, pp. 11-19, 2021.
[10] A. A. Kresna, "The Epistemology of Rasa as a Basic Foundation of the Javanese Psychology," *East Asian Journal of Multidisciplinary Research*, vol. 2, no. 8, pp. 3209-3222, 2023.
[11] T. A. R. Yunanto, "Happiness in the Javanese context: Exploring the role of emotion regulation and resilience," *Humanitas: Indonesian Psychological Journal*, pp. 149-158, 2023.
[12] Y. Khurana, S. Gupta, R. Sathyaraj, and S. Raja, "RobinNet: A Multimodal Speech Emotion Recognition System With Speaker Recognition for Social Interactions," *IEEE Transactions on Computational Social Systems*, 2022.
[13] G. Wen, S. Ye, H. Li, P. Wen, and Y. Zhang, "Multimodal and Multitask Learning with Additive Angular Penalty Focus Loss for Speech Emotion Recognition," *International Journal of Intelligent Systems*, vol. 2023, no. 1, p. 3662839, 2023.
[14] R. A. Patamia, P. E. Santos, K. N. Acheampong, F. Ekong, K. Sarpong, and S. Kun, "Multimodal Speech Emotion Recognition Using Modality-Specific Self-Supervised Frameworks," in *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2023: IEEE, pp. 4134-4141.
[15] G.-N. Dong, C.-M. Pun, and Z. Zhang, "Temporal relation inference network for multimodal speech emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6472-6485, 2022.
[16] 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 (IJEEI)*, vol. 10, no. 3, pp. 584-591, 2022.
[17] T. Ahmed, I. Begum, M. S. Mia, and W. Tasnim, "Multimodal Speech Emotion Recognition Using Deep Learning and the Impact of Data Balancing," in *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, 2023: IEEE, pp. 1-6.
[18] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," *Entropy*, vol. 25, no. 10, p. 1440, 2023.
[19] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 73-79, 2021.
[20] K. Nugroho, E. Noersasongko, and H. A. Santoso, "Javanese gender speech recognition using deep learning and singular value decomposition," in *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2019: IEEE, pp. 251-254.

BIOGRAPHY OF AUTHORS

Dr. Fatchul Arifin, a distinguished academic and engineer, began his illustrious educational journey at Universitas Diponegoro, earning a B.Sc. in Electrical Engineering in 1996. He furthered his studies at the prestigious Institut Teknologi Bandung (ITB), obtaining a Master's degree in Electrical Engineering in 2003. His quest for knowledge led him to Institut Teknologi Surabaya, where he completed his Doctoral degree in Electrical Engineering in 2014. In 2017, Dr. Arifin achieved his professional engineering title from Universitas Negeri Yogyakarta (UNY).

He is an esteemed lecturer at UNY, where he teaches undergraduate and postgraduate courses in the Electronic and Informatic programs of the Engineering Faculty. His research interests are broad and interdisciplinary, spanning Artificial Intelligence, Machine Learning, Fuzzy Logic, and Biomedical Engineering. He has held significant academic leadership roles, including serving as the head of the Department of Electronic and Informatic Engineering Education and as the Chairman of the Central Java and Yogyakarta Regional Forum for Electrical Engineering (FORTEI). He is currently the Head of the Master's Program in Electronic and Informatic Engineering at UNY.

Aris Nasuha received a B.Sc. in Physics at Universitas Gadjah Mada in 1993. Afterward, he received a Master's and Doctoral degree from the Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember Surabaya (ITS) in 2001 and 2019. His research interests include but are not limited to Artificial Intelligent Systems, Machine Learning, and Digital Signal Processing. He is a lecturer at the Department of Electronics and Informatic Engineering (1994-now) and head of Bachelor Applied Science in Electronic Engineering (2019-now) at the Faculty of Engineering, Universitas Negeri Yogyakarta.

Ardy Seto Priambodo obtained his B.Eng. degree in Electrical Engineering from Institut Teknologi Sepuluh Nopember (ITS), Indonesia, 2012. He received his M.Eng. degree 2018 from Electrical Engineering, Universitas Gadjah Mada (UGM), Indonesia. He obtained a professional engineering degree (Ir.) from Universitas Negeri Yogyakarta in 2020. His research interests are control systems, instrumentation, and robotics. He is currently a lecturer at Universitas Negeri Yogyakarta (UNY), Indonesia (2019-now), Head of Instrumentation Laboratory (2020-now), and Flying Robot UNY Team Supervisor (2019-now).

Anggun Winursito is a researcher focused on signal processing and instrumentation. He received an M.Eng. from Gadjah Mada University and a B.Ed. from Yogyakarta State University in 2014 and 2018. His areas of expertise are speech recognition, image recognition, analog and digital signal processing, and instrumentation. He has participated in several local and national research projects, including speech pattern recognition and radar telecommunications systems. He was awarded the best paper at an international conference held in Malaysia for his research on the compression of speech features to improve the accuracy of speech recognition systems. Currently, he is a lecturer at the Department of Electronic and Informatic Engineering Education, Faculty of Engineering, Universitas Negeri Yogyakarta.

Teddy Surya Gunawan (Senior Member, IEEE) received the B.Eng. degree (cum laude) in electrical engineering from the Institut Teknologi Bandung (ITB), Indonesia, in 1998, the M.Eng. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2001, and the Ph.D. degree from the School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia, in 2007. His research interests include speech and audio processing, biomedical signal processing and instrumentation, image and video processing, and parallel computing. He was awarded the Best Researcher Award from IIUM in 2018. He was a Chairman of the IEEE Instrumentation and Measurement Society–Malaysia Section (2013, 2014, and 2020), a Professor (since 2019), the Head of Department (from 2015 to 2016) with the Department of Electrical and Computer Engineering, and the Head of Programme Accreditation, and the Quality Assurance for Faculty of Engineering (from 2017 to 2018), International Islamic University Malaysia. He has been a Chartered Engineer (IET, U.K.) and Insinyur Profesional Madya (PII, Indonesia) since 2016, a registered ASEAN Engineer since 2018, and an ASEAN Chartered Professional Engineer since 2020.