Improving Emotion Recognition Accuracy Using a Multimodal Model (Face and Voice Video) Based on a Convolutional Neural Network (CNN)

Karnadi[1], Ermatita[2]*, Abdiansah[3]
Doctor of Engineering Program, Universitas Sriwijaya Palembang, Indonesia[1]
Faculty of Computer Science, Universitas Sriwijaya Palembang, Indonesia[2, 3]

***Abstract*—Advancements in Artificial Intelligence (AI) technology have enabled the recognition of human emotions. Along with the development of deep learning and multimodal processing methods, emotion analysis can now be performed by utilizing multiple data sources simultaneously, such as facial expressions and speech signals. However, existing emotion recognition systems still face limitations in terms of accuracy. This study aims to develop and evaluate a more accurate emotion recognition system by implementing a Convolutional Neural Network (CNN)-based prediction model that integrates facial and audio data simultaneously. The study utilizes the CREMA-D dataset, which consists of visual data in the form of facial images and audio data containing variations of emotional expressions. The research process includes data preprocessing, feature extraction, and multimodal integration using an optimized Convolutional Neural Network (CNN) architecture. The evaluation results based on the F1-score indicate that the multimodal facial and audio data enable the model to recognize emotions effectively. Model performance was measured using accuracy, precision, recall, and F1-score metrics. Experimental results show that the angry (ANG) class achieved the best performance with an F1-score of 82%, while the fear (FEA) class demonstrated the lowest performance with an F1-score of only 58%. The results further indicate that the multimodal model achieved higher accuracy than unimodal models, significantly improving generalization capability on diverse testing data. This study demonstrates an overall emotion recognition accuracy improvement of 69% through the combination of facial and audio features. The analysis of combined facial and speech features on emotion classification performance shows that the proposed model achieves good overall performance, where the integration of image and audio modalities improves the correctness of facial expression predictions. Future research is expected to further improve accuracy by incorporating additional modalities beyond facial and audio data.**

***Keywords*—Emotion recognition; CNN; multimodal learning; CREMA-D; face; voice; accuracy; deep learning**

I. Introduction

An important factor in human communication is emotion, [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], which influence behavior, perception, and decision-making. Research on automatic human emotion recognition has advanced rapidly thanks to advances in artificial intelligence, particularly the development of systems capable of understanding and responding to users' emotional states in a more human-like manner. To recognize human emotions more accurately, research has utilized single-modal observations [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] of either voice or facial data. Previous researchers [29], [30], [31], [32], [33], used the VGG Face architecture to extract facial features and MFCC (Mel-Frequency Cepstral Coefficients) for voice features. This system can be used to help develop more user-friendly and responsive human-computer interfaces (HCI) [34]. A three-layer system was developed, consisting of FEAM (facial analysis) [35], [36], [37], VSAM (voice analysis) [38], [39], [40], [41], [42], [43], and MERM (emotion merging for the metaverse) [44], [45], [46]. This system is designed to enhance the realism and accuracy of emotional interactions in virtual environments such as the metaverse. This system can identify seven primary emotions in real time by utilizing technologies such as FCNN (Fully Convolutional Neural Network) and MCycleGAN (Modified CycleGAN) [47], [48], [49]. In simulation experiments, the results were comparable to human perceptual capabilities, focusing on the use of voice and facial recognition technologies for social inclusion [50], [51], [52], [53], [54], [55], [56]. This system uses OpenCV, a Raspberry Pi, and facial recognition techniques such as Haar cascades and LBPH, as well as voice control, to open automatic doors. Test results show that this system is highly effective in helping people with disabilities access their homes independently, with an accuracy rate of 99.63%. All three studies indicate that the integration of auditory and visual data holds great potential for helping to develop intelligent systems that are responsive to real human needs [57]. The uncertainty that arises when using a single modality is one of the main challenges in emotion recognition. Audio signals are affected by ambient noise and the quality of the recording device, while lighting, the camera angle, or even deliberate facial expressions can influence the recording. Multimodal methods leverage the strengths of each modality and mitigate their weaknesses through the process of information fusion; this method offers a solution. In this study, we used CREMA-D, a crowdsourced multimodal dataset. This dataset consists of video recordings of actors expressing six basic emotions: happiness, sadness, anger, fear, frustration, and rage. This dataset is ideal for multimodal emotion recognition research because it includes visual, audio, and audiovisual data. To extract spatiotemporal features of facial expressions from the video data, a CNN-LSTM architecture was used. Meanwhile, the audio data was converted into Mel-spectrograms for feature extraction using a Convolutional Neural Network (CNN).

*Corresponding author.

Multimodal emotion recognition (MER) has emerged as a promising approach in this context because it can integrate various sources of emotional information, such as facial expressions and voice intonation, into a single approach. Research findings indicate that this multimodal method outperforms single-modal methods in detecting emotions.

Given these opportunities and challenges, this study aims to develop a multimodal model based on a Convolutional Neural Network (CNN) capable of integrating visual and audio information to improve the accuracy of emotion recognition.

II. Materials and Methods

A. Multimodal

Multimodal machine learning (also known as multimodal learning) is a subfield of machine learning that aims to develop and train models capable of utilizing various types of data and learning to link or combine these modalities, intending to improve predictive performance [58]. Deep learning (DL), as a cutting-edge technology, has achieved remarkable breakthroughs in many computer vision tasks due to its impressive capabilities in data representation and reconstruction. Naturally, it has been successfully applied to the field of multimodal hospital data fusion, yielding significant improvements over traditional methods. This survey aims to provide a systematic overview of deep learning-based multimodal hospital data fusion. More specifically, some key insights on this topic are presented first [59]. According to researchers [60], [61], [62], six machine learning models employing two distinct feature extraction techniques are considered foundational models.

B. Convolutional Neural Network Algorithm

A convolutional neural network is a type of neural network designed to process grid-based data, such as two-dimensional images. The term "convolution" refers to a linear algebraic operation that multiplies a filter matrix by the image to be processed [63], [64], [65], [66]. Fig. 1 shows the architecture of the proposed CNN model for facial expression recognition.

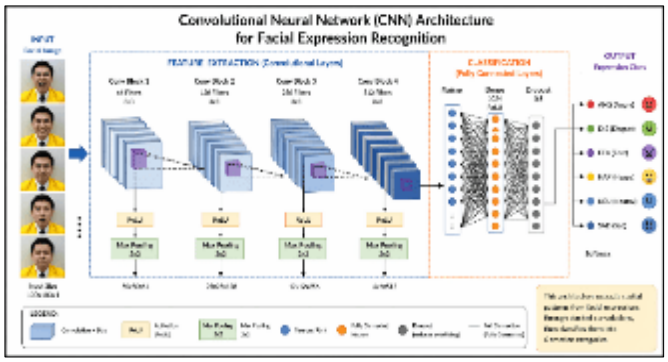


Fig. 1. Convolutional neural network architecture.

1) Data split (stratified split): Data splitting (stratified split) according to [66]. The Scikit-learn function used to split a dataset into two parts—training data and test data—is written in the Python code `train_test_split(...)`. In general, the splitting of a dataset into training and test data [67], [68], [69], [70], [71], can be expressed as follows:

$$D = \{(x_i, a_i, y_i)\}_{i=1}^{N} \quad (1)$$

$With: x_i, \epsilon\ X_{img}, a_i \in\ X_{audio}, y_i \in Y$

Divide D by:

$$D_{train} = \{(x_i, a_i, y_i)\}_{i\ =\ 1}^{0.8\ N}$$

$$D_{test} = \{(x_j, a_j, y_j)\}_{j\ =\ 1}^{0.8\ N}$$

$by\ stratification: P(y_i \in\ c_k)\ on\ D_{train} \approx P(y_i \in c_k)\ on\ D, \forall_{c_k} \in\ class$

2) Cutting to fit the batch: This function is used to reduce the amount of data by a multiple of the batch size. This is particularly important to avoid having an incomplete final batch, especially in models with multimodal inputs, such as images and audio recordings [72], [73], [74], [75].

The general formula is as follows:

$$n_{batch} = \left\lfloor \frac{N}{B} \right\rfloor \quad (2)$$

$$N' = n_{batch} . B$$

$$D' = \left\{ \left(x_i^{(1)}, x_i^{(2)}, y_i\right) \right\}_{i\ =\ 1}^{N'}$$

3) Training model: This formula is used to evaluate model performance during both training and testing [75], [76], [77]. The loss function plays a crucial role in the backpropagation process for updating the weights of a neural network through optimization (such as gradient descent) [78], [79]. It implicitly uses that average formula behind the scenes. Here is the formula used:

$$Loss = \frac{1}{N} \sum_{i=1}^{N} L(\hat{y}i, yi) \quad (3)$$

4) Predictions and evaluations: Predictions and evaluations are performed using the final-class prediction formula in a classification model that employs the softmax activation function. This formula is the standard formula used in multiclass classification with deep neural networks, particularly in the output layer of classification models that use the softmax activation function [80], [81], [82]. Here is the formula used:

$$y^{\wedge} = \text{argmax}(\text{softmax}(z)) \quad (4)$$

Proposed approach

C. Research Methods and Types

The objective of this experimental quantitative study, which employs a deep learning-based systems engineering approach, is to develop and evaluate a multimodal emotion recognition system using Convolutional Neural Networks (CNN).

D. Data Sources and Data Types

This study uses secondary data in the form of the CREMA-D dataset (Crowd-sourced Emotional Multimodal Actors Dataset), which is a publicly available multimodal dataset

accessible via the Kaggle platform (https://www.kaggle.com/datasets/orvile/crema-d-emotional-multimodal-dataset). The CREMA-D dataset consists of 7,442 video clips featuring 91 actors (48 men and 43 women) aged 20–74 who express six basic emotion categories: anger, happiness, sadness, fear, disgust, and neutral. Each sample in the dataset has been manually labeled and contains two main modalities: visual data in the form of facial expressions recorded in video format, and audio data in the form of voice intonation that can be extracted into acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC). This dataset was selected for its high annotation quality and good inter-rater reliability, its representative demographic diversity of actors, and its multimodal format, which aligns with the research objectives of developing a CNN model that integrates visual and audio information for emotion regulation identification.

E. Research Phases

The research process consists of several stages carried out by the author, including: data collection, data preprocessing, multimodal feature fusion, model training, model evaluation, and system implementation, as shown in Fig. 2 below [83], [84].

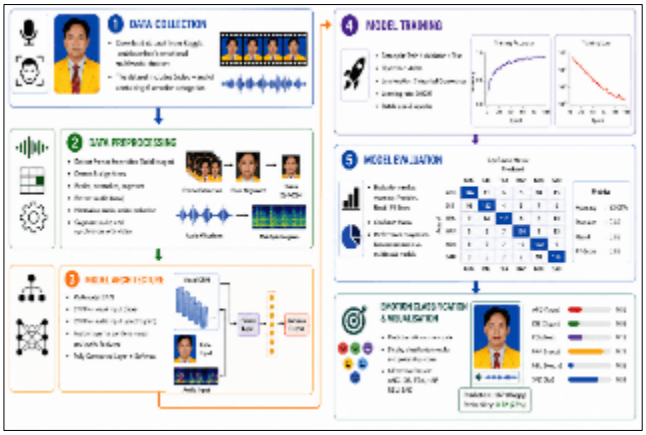


Fig. 2. Research phases.

The data collection phase involved downloading the CREMA-D dataset from the Kaggle platform, which contains 7,442 multimodal video clips across six emotion categories (anger, happiness, sadness, fear, disgust, neutral). This dataset was selected because it provides two modalities—facial expressions (visual) and voice intonation (audio)—both of which are well-labeled. The preprocessing stage involves extracting data from video files into two separate components. For the visual modality, face detection, alignment, cropping, and resizing of facial images to 64×64 pixels are performed. For the audio modality, MFCC (Mel-Frequency Cepstral Coefficients) features are extracted and converted into a spectrogram with dimensions of 40×128. Data augmentation is used to increase the diversity of the dataset and prevent overfitting.

This stage designs a dual-stream-based multimodal CNN architecture consisting of an Image Branch for processing facial images and an Audio Branch for processing audio spectrograms. Both branches use three convolutional blocks (32, 64, 128 filters) with Batch Normalization and Max Pooling. Features from both modalities are combined using concatenation in the Fusion Layer, then processed through fully connected layers for final classification.

The model was trained using a dataset split into a stratified split (70:15:15 for training, validation, and testing). The Adam optimizer with a learning rate of 0.001, a categorical cross-entropy loss function, and early stopping was used to optimize the network parameters. Training was conducted for up to 100 epochs with a batch size of 32 until the model reached convergence.

Model performance was evaluated using test data and various metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. A comparison was conducted between the multimodal model and unimodal models (visual-only and audio-only) to assess the effectiveness of the fusion strategy. Statistical analysis was performed to validate the significance of the performance improvement.

The final stage presents visualizations of the classification results, including a confusion matrix, graphs showing training and validation loss and accuracy, and examples of predictions on the test set. The visualizations also include Grad-CAM to interpret the facial regions that most influence the predictions and feature importance analysis to understand the contribution of each modality to the model's classification decisions.

F. CNN Model Architecture

The proposed architecture employs a dual-stream Convolutional Neural Network (CNN) approach with two separate branches to independently process visual and audio modalities. The visual branch is designed to extract facial expression features, while the audio branch utilizes a 40×128 MFCC spectrogram as the representation of speech features. The extracted feature representations from both modalities are subsequently fused through a concatenation mechanism at the fusion layer to construct an integrated multimodal representation before the emotion classification process. Furthermore, the dataset was divided using a stratified split strategy with a ratio of 70:15:15 for training, validation, and testing sets, respectively, in order to preserve class distribution balance and ensure a reproducible experimental setup. The emotion classification process using these results is shown in Fig. 3 below [85], [86], [87], [88], [89], [90], [91], [92], [93], [94].

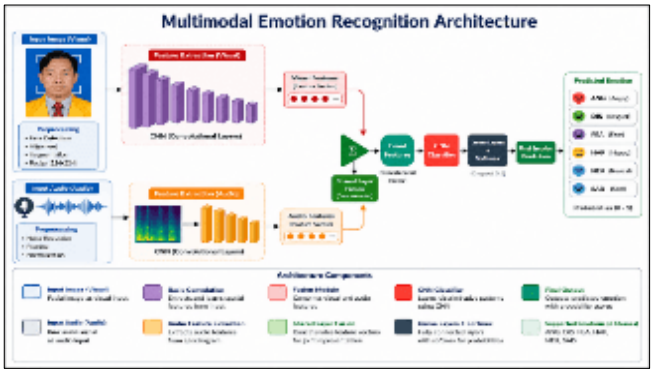


Fig. 3. CNN architecture.

Fig. 3 above shows the model architecture, which consists of two parallel branches for processing visual and audio modalities independently. In the Image Branch, 64×64×3 image inputs are

processed through three convolutional blocks with progressively larger filters (32, 64, 128), each equipped with Batch Normalization and Max Pooling. Global Average Pooling then aggregates the feature maps, followed by a 128-neuron Dense layer with Dropout, producing a 128-dimensional image feature vector.

In the Audio Branch, a 40×128×1 audio spectrogram is processed using a CNN architecture with the same filter configuration (32, 64, 128). Following Global Average Pooling, a dense layer with 128 neurons and Dropout produces a 128-dimensional audio feature vector that encodes the prosodic and spectral characteristics of the audio.

The Fusion Layer combines the two feature vectors via concatenation, producing a 256-dimensional multimodal representation. This representation is processed through two Dense layers (with 256 and 128 neurons) using Batch Normalization and Dropout to learn complex interactions between modalities. The Output Layer uses a Dense layer with a Softmax activation function to generate a probability distribution for each emotion regulation class, enabling the model to learn optimal representations and fusion strategies end-to-end.

TABLE I. DUAL STREAM CNN ARCHITECTURE FOR MULTIMODAL EMOTION RECOGNITION

Layer	Output Shape	Parameter
Input Image	255	238
Conv2D + ReLU	255	271
MaxPooling2D	255	145
Conv2D + ReLU	255	175
MaxPooling2D	215	229
MaxPooling2D	(32,32,64)	Pool size 2×2
GlobalAveragePooling	(64)	-
Input Audio MFCC	(40,128,1)	MFCC spectrogram
Conv2D + ReLU	(40,128,32)	Kernel 3×3
MaxPooling2D	(20,64,32)	Pool size 2×2
Conv2D + ReLU	(20,64,64)	Kernel 3×3
MaxPooling2D	(10,32,64)	Pool size 2×2
GlobalAveragePooling	(64)	-
Concatenate	(128)	Feature fusion
Dense + ReLU	(128)	Fully connected
Dropout	(128)	Rate = 0.5
Output Layer	(6)	Softmax activation

Table I presents the proposed dual-stream CNN architecture for multimodal emotion recognition, which consists of separate visual and audio processing branches. The visual branch processes RGB facial images through multiple convolutional and max-pooling layers to extract facial expression features, while the audio branch processes MFCC spectrogram inputs using a similar CNN structure to extract acoustic features. The extracted features from both modalities are then combined using a concatenation layer for multimodal feature fusion. The fused representation is further processed through fully connected and dropout layers before being classified into six emotion classes using a Softmax output layer. This architecture is designed to effectively learn complementary information from facial expressions and speech signals to improve emotion recognition performance.

G. Hardware and Software

Programming Language Tools and Software: Python Framework: TensorFlow/Keras or PyTorch Hardware: Nitro ANV16-71 Laptop with Intel(R) Core(TM) i7-14650HX processor, GPU, and 16 GB of RAM Supporting Software: OpenCV, Librosa, and Scikit-learn.

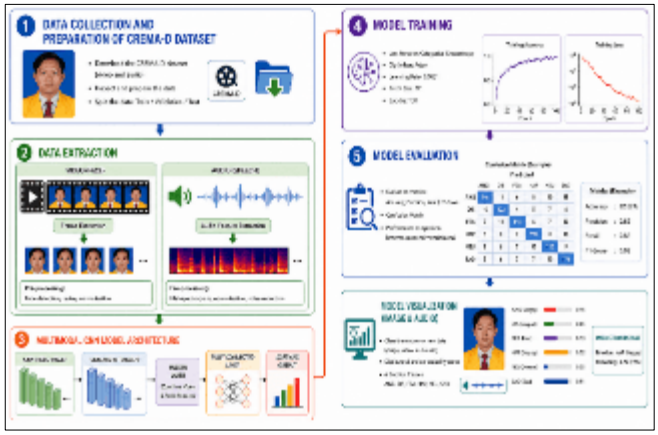


Fig. 4. Flowchart of the stages of multimodal deep learning analysis using the CNN algorithm.

Fig. 4 above illustrates the workflow of the multimodal emotion recognition system. The process begins with data collection, which includes video and audio. Both types of data then proceed to the preprocessing stage: video preprocessing to process facial frames and audio preprocessing to process audio signals. After that, the video and audio are processed by the CNN Feature Extract module to extract key features from each modality. The resulting features are then combined in the feature fusion stage, allowing the model to obtain a more comprehensive representation of emotions. Finally, the combined results are classified in the emotion classification stage to produce emotion categories.

III. RESULTS AND DISCUSSION

A. Data Collection

The dataset used in this study is the CREMA-D dataset. This dataset contains videos of actors expressing six types of emotions: happy, sad, angry, afraid, annoyed, and neutral. This dataset supports the development of voice- and image-based emotion recognition systems. To speed up the model training process, the dataset was processed using the Google platform

- Optimizer: Adam, Loss: Categorical Cross-entropy or Sparse Categorical.
- Regularization techniques: Dropout, EarlyStopping, ReduceLROnPlateau.

- Batch size and training time are adjusted based on GPU capacity; use Colab with GPU access after downloading from Kaggle.

B. Preprocessing

The datasets obtained in Section A were preprocessed. The preprocessing steps for each dataset are as follows: CREMA-D: OpenCV and a Haar cascade were used to extract faces from videos; the face images were resized to 64 x 64 pixels (for CNN) or 224 x 224 pixels (for the customized model). Audio is extracted using FFmpeg and converted into 40 x 128-dimensional MFCCs (Mel-Frequency Cepstral Coefficients). Normalization is performed to maintain input consistency.

C. Multimodal Feature Fusion

Multimodal feature fusion was performed by combining feature representations from visual and audio modalities to develop a multimodal emotion recognition system. In the visual branch, facial images were processed using a 2D CNN architecture to extract facial expression features, while in the audio branch, speech signals from the CREMA-D dataset were converted into 40×128 MFCC spectrograms and processed using a 2D CNN architecture to extract acoustic features. The extracted features from both modalities were subsequently combined at the feature fusion layer using the concatenate () method to form an integrated multimodal representation. The fused representation was then forwarded to the fully connected layer for the final emotion classification process.

D. Model Training

In this model training, the data used consists of facial and voice data obtained from the CREMA-D dataset. A multimodal Convolutional Neural Network (CNN) model was trained to classify human emotions based on the combination of facial expressions and voice features.

Before the training process was carried out, the initial step involved determining the storage location of the dataset to be used. The CREMA-D dataset was successfully uploaded and extracted into the Google Colab workspace, allowing all data to be accessed and processed by the system during the model training stage. The folder structure used consists of the main CREMA-D directory containing video data, audio data, and other supporting information required for the emotion classification process. The folder organization of the CREMA-D dataset includes the following directories:

- AudioMP3: These are .mp3 audio files that record the actor's vocal expressions based on specific emotions.
- AudioWAV: These contain the same audio files as AudioMP3 but in .wav format. This format is often used for audio signal processing because it is uncompressed and preserves the original quality.
- VideoFlash folder: This contains videos in .flv format that record both the actors' facial expressions and voices simultaneously. This allows for use in research involving the integration of visual and audio emotions.
- Docs folder: The docs folder may contain instructions or additional information about the dataset. This may include explanations of the file naming structure, emotion annotations, usage guidelines, and so on.
- ProcessedResults folder: The processedResults folder typically contains the results of processing or extracting raw data, such as feature extraction results, processed annotations, or format conversions.
- License.txt: The text file named License.txt contains the official license for the CREMA-D dataset. It is important to understand the restrictions on data use in research or publications.

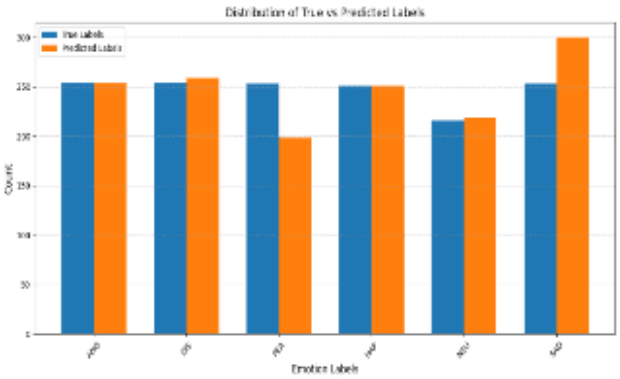


Fig. 5. Actual label distribution vs. model predictions on emotion data.

Fig. 5, the "True vs Predicted Distribution" chart, presents a comparison between the actual labels (True Labels) and the model predictions (Predicted Labels) across six emotion classes: ANG, DIS, FEA, HAP, NEU, and SAD. Overall, the true label distribution is relatively balanced, with most classes containing approximately 250 samples, while the NEU class has a slightly lower number of samples. The predicted distribution closely matches the true distribution for the ANG, DIS, HAP, and NEU classes, indicating that the model performs relatively well in recognizing these emotions. However, a noticeable discrepancy can be observed in the FEA class, where the number of predicted samples is significantly lower than the actual distribution, suggesting underprediction for this emotion category. In contrast, the SAD class shows a considerably higher number of predicted samples compared to the true labels, indicating that the model tends to over-predict the SAD emotion. Overall, the chart demonstrates that the model is capable of capturing the general distribution of most emotion classes, although some prediction bias remains, particularly between the FEA and SAD classes.

TABLE II. ACTUAL AND PREDICTED NUMBER OF LABELS BY EMOTION CLASS

Emotions Class	Number of Actual Data Points (True Labels)	Number of Model Predictions (Predicted Labels)	Difference (Predicted vs. Actual)
ANG (Angry)	255	255	0
DIS (Disgust)	255	260	+5
FEA (Fear)	255	200	-55
HAP (Happy)	250	250	0
NEU (Neutral)	215	220	+5
SAD (Sad)	255	300	+45

Table II presents the comparison between the number of actual data points (true labels) and the number of model predictions (predicted labels) for each emotion class. Overall, the model demonstrates relatively balanced prediction performance for several classes, particularly ANG and HAP, where the number of predicted samples exactly matches the actual data distribution. The DIS and NEU classes show only minor differences, with slight over-predictions of +5 samples each, indicating stable classification performance. However, significant discrepancies can be observed in the FEA and SAD classes. The FEA class is notably underpredicted by 55 samples, suggesting that the model has difficulty correctly identifying fear-related emotions. In contrast, the SAD class is over-predicted by 45 samples, indicating a tendency of the model to classify several other emotions as sadness. These results suggest that while the model is capable of maintaining good prediction consistency for some emotion categories, challenges still remain in distinguishing emotions with similar facial and acoustic characteristics, particularly between FEA and SAD.

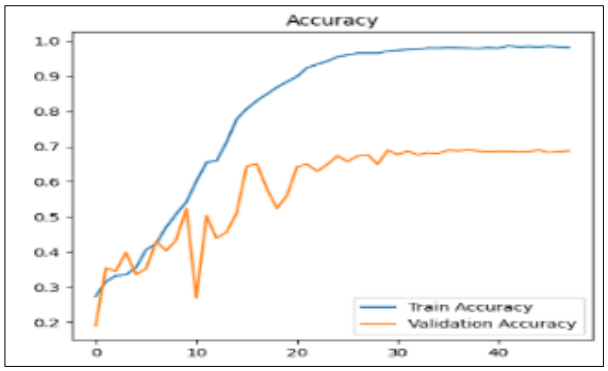


Fig. 6. Accuracy over epochs.

Fig. 6 illustrates the training and validation accuracy curves of the proposed multimodal CNN model during the training process. The training accuracy shows a consistent upward trend, eventually reaching approximately 98%, indicating that the model is capable of learning the training data effectively. Meanwhile, the validation accuracy also improves over time and stabilizes around 68–69%, although several fluctuations can be observed during the early and middle training epochs. The gap between training and validation accuracy suggests the presence of moderate overfitting; however, the validation performance remains relatively stable in the later epochs, indicating that the model still maintains reasonable generalization capability on unseen data.

Fig. 7 illustrates the training loss and validation loss curves of the proposed multimodal CNN model during the training process. The training loss decreases consistently and approaches zero, indicating that the model is able to learn the training data effectively. Meanwhile, the validation loss initially decreases but later exhibits noticeable fluctuations and tends to stabilize around 1.3–1.5 during the final epochs. The considerable gap between training loss and validation loss suggests the presence of overfitting, where the model performs very well on the training data but shows limited performance on the validation data. Nevertheless, the relatively stable validation loss in the later epochs indicates that the model still maintains a reasonable generalization capability on unseen data.

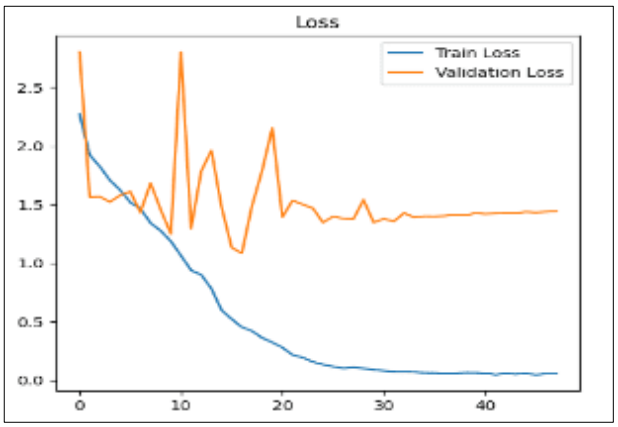
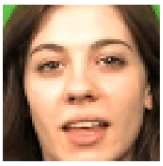
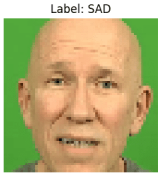


Fig. 7. Loss over epochs.

The video file format used in this study is .flv (Flash Video). This can be observed from the following code snippet: video_files = [f for f in os.listdir(video_dir) if f.endswith('.flv')], which indicates that the system only processes video files with the .flv extension. The .flv format was selected because it has a relatively small file size and is compatible with various video processing applications, thereby supporting efficiency in the video extraction and analysis processes in this study. Examples of facial and audio prediction results generated by the model used in this study are presented in Table III.

TABLE III. EXAMPLE OF SIX FACES AND AUDIO IMAGE PREDICTION

Label: NEU  Sample ke-1 — 1003_ITH_NEU_XX.flv	Label: FEA  Sample ke-2 — 1036_TSI_FEA_XX.flv	Label: HAP  Sample ke-3 — 1012_WSI_HAP_XX.flv
Label: DIS  Sample ke-4 — 1041_IOM_DIS_XX.flv	Label: SAD  Sample ke-5 — 1062_DFA_SAD_XX.flv	Label: DIS  Sample ke-6 — 1076_IWL_DIS_XX.flv

As shown in Table III, the model is capable of predicting emotions based on the combination of facial and audio data from several video samples.

The classification report generated from the evaluation of the multimodal (face and audio) emotion classification model is shown below. This report includes key evaluation metrics for each emotion class tested—namely, accuracy, recall, and F1-score as well as a summary of the model's overall performance.

Table IV presents the classification performance across six emotion classes using precision, recall, F1-score, and support (number of samples per class). The model performs best on the ANG class, achieving the highest F1-score of 0.82 with strong precision (0.82) and recall (0.82), indicating reliable and

balanced predictions. The HAP class also shows relatively good performance with an F1-score of 0.81. In contrast, DIS and FEA have the lowest F1-scores (0.67 and 0.58), reflecting the model's difficulty in accurately distinguishing these emotions. The SAD class exhibits high recall (0.61) but lower precision (0.66), suggesting that while many SAD instances are correctly identified, the model tends to overpredict this class. The NEU class shows moderate and balanced performance with an F1-score of 0.66. Overall, the model achieves an accuracy of 0.69 across 1,481 samples, with both macro and weighted averages also at 0.69, indicating consistent but still moderate performance and highlighting the need for further improvements in distinguishing similar emotional expressions.

TABLE IV. CLASSIFICATION REPORT

	Precision	Recall	F1-Score	support
ANG	0.82	0.82	0.82	254
DIS	0.67	0.68	0.67	254
FEA	0.66	0.51	0.58	253
HAP	0.81	0.81	0.81	251
NEU	0.66	0.67	0.66	216
SAD	0.56	0.66	0.61	253
Accuracy			0.69	1481
Macro avg	0.69	0.69	0.69	1481
Weighted avg	0.70	0.69	0.69	1481

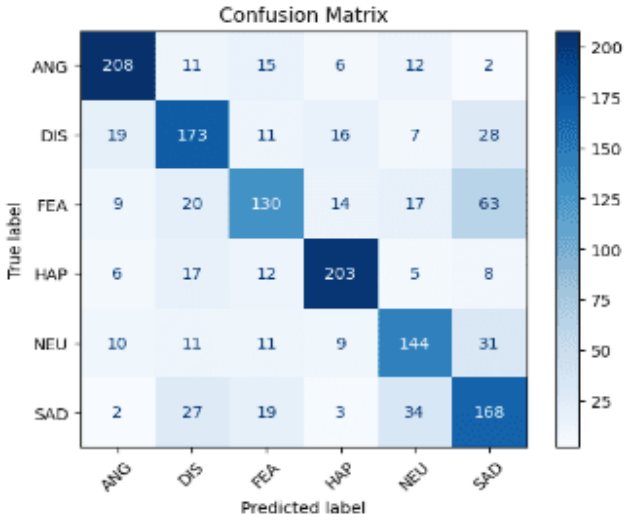


Fig. 8. Confusion matrix of emotion classification results using a multimodal CNN model.

Fig. 8 shows the confusion matrix illustrating the performance of the emotion classification model across six emotion classes, namely ANG, DIS, FEA, HAP, NEU, and SAD, where correct predictions are represented along the diagonal of the matrix. The model demonstrates the best performance on the ANG class with 208 correct predictions, followed by HAP with 203 correct predictions, DIS with 173 correct predictions, and SAD with 168 correct predictions, indicating strong recognition capability for these emotion categories. The NEU class also shows relatively good performance with 144 correctly classified samples, while the FEA class has the lowest performance with 130 correct predictions. Overall, the confusion matrix indicates that the proposed model achieves good classification performance across most emotion categories; however, it still faces challenges in distinguishing several negative emotions with similar characteristics, leading to misclassifications between classes. Visualization of Multimodal Emotion Classification Results with Confidence Scores.

TABLE V. VISUALIZATION OF MULTIMODAL EMOTION CLASSIFICATION RESULTS WITH CONFIDENCE SCORES

Class Label	Class Acc (%)	Audio Means (%)	Confidence (%)
ANG 	81.89	-0.69	92.48
DIS 	68.11	-0.17	99.91
FEA 	51.38	-0.26	83.57
HAP 	80.88	-0.58	99.99
NEU 	0.66	0.67	0.66
SAD 	0.56	0.66	0.61

Table V presents the visualization results of facial emotion classification using the proposed multimodal CNN model across six emotion classes, namely angry (ANG), disgust (DIS), fear (FEA), happy (HAP), neutral (NEU), and sad (SAD). Each facial sample is accompanied by class accuracy, audio mean,

and confidence score values, which indicate the model's confidence level in performing the prediction. The model achieves the best performance on the ANG class with an accuracy of 81.89% and the HAP class with an accuracy of 80.88%, while the FEA class shows the lowest performance with an accuracy of 51.38%, indicating that fear emotion remains difficult to distinguish from other emotions. The confidence scores for most classes are above 90%, demonstrating that the model has a high level of confidence in its classification results. In addition, the audio mean values represent the average contribution of audio features for each emotion prediction. Overall, this visualization demonstrates that the multimodal model is capable of recognizing most emotions effectively through the combination of visual and audio information, although challenges still remain in classifying emotions with similar facial expression characteristics.

IV. Conclusion

The results of this study demonstrate that the proposed multimodal CNN model is capable of effectively recognizing human emotions by integrating facial image and audio features. Based on the evaluation results, the model achieved an overall accuracy of 69%, with weighted precision, recall, and F1-score values of 0.70, 0.69, and 0.69, respectively. Among the evaluated emotion classes, the ANG (Angry) class achieved the best performance with an F1-score of 0.82, followed by the HAP (Happy) class with an F1-score of 0.81, indicating that the model was highly effective in recognizing these emotions. In contrast, the FEA (Fear) class showed the lowest performance with an F1-score of 0.58, suggesting that fear-related emotions remain challenging to distinguish due to similarities with other negative emotional expressions. The experimental results further indicate that the proposed multimodal approach was able to integrate visual and audio information effectively, resulting in stable performance across several major emotion classes. In addition, the developed CNN architecture provides an efficient and practical framework for multimodal emotion recognition through feature fusion between facial expressions and speech signals. Overall, this study confirms that combining facial and audio modalities can improve emotion classification performance and enhance the model's generalization capability on diverse testing data. Future work may focus on improving recognition performance for challenging emotion classes and incorporating additional modalities or advanced fusion strategies to further enhance classification accuracy.

Acknowledgment

The authors would like to express their sincere gratitude to all parties who have provided support, assistance, and contributions throughout the completion of this research. Special appreciation is extended to the authors and their families for their prayers, motivation, and moral support. The authors also gratefully acknowledge Universitas Sriwijaya, Palembang, for providing academic facilities, a supportive research environment, and institutional support that made this study possible.

References

[1] C. E. Valderrama, M. G. Gomes Ferreira, J. M. Mayor Torres, A. R. Garcia-Ramirez, and S. G. Camorlinga, "Editorial: Machine learning approaches to recognize human emotions," Front. Psychol., vol. 14, 2023, doi: 10.3389/fpsyg. 2023.1333794.

[2] C. Qiana, J. A. L. Marquesb, and A. R. de Alexandria, "Real-time emotion recognition based on facial expressions using Artificial Intelligence techniques: A review and future directions," Multidiscip. Rev., vol. 8, no. 10, pp. 1–27, 2025, doi: 10.31893/multirev. 2025328.

[3] Y. Wu, Q. Mi, and T. Gao, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions," 2025. doi: 10.3390/biomimetics10070418.

[4] R. Pereira et al., "Systematic Review of Emotion Detection with Computer Vision and Deep Learning," 2024. doi: 10.3390/s24113484.

[5] T. Chutia and N. Baruah, "A review on emotion detection by using deep learning techniques," Artif. Intell. Rev., vol. 57, no. 8, p. 203, 2024, doi: 10.1007/s10462-024-10831-1.

[6] M. Karthiga, E. Suganya, S. Sountharrajan, B. Balusamy, and S. Selvarajan, "EEG-based smart emotion recognition using meta heuristic optimization and hybrid deep learning techniques," Sci. Rep., vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-80448-5.

[7] X. Wang, Y. Ren, Z. Luo, W. He, J. Hong, and Y. Huang, "Deep learning-based EEG emotion recognition: Current trends and future perspectives.," Front. Psychol., vol. 14, p. 1126994, 2023, doi: 10.3389/fpsyg. 2023.1126994.

[8] E. Gkintoni, A. Aroutzidis, H. Antonopoulou, and C. Halkiopoulos, "From Neural Networks to Emotional Networks: A Systematic Review of EEG-Based Emotion Recognition in Cognitive Neuroscience and Real-World Applications.," Brain Sci., vol. 15, no. 3, Feb. 2025, doi: 10.3390/brainsci15030220.

[9] A. M. S. Gonzalez-Acosta et al., "The first look: a biometric analysis of emotion recognition using key facial features," Front. Comput. Sci., vol. 7, 2025, doi: 10.3389/fcomp. 2025.1554320.

[10] E. S. Agung, A. P. Rifai, and T. Wijayanto, "Image-based facial emotion recognition using a convolutional neural network on the Emotion Detection Dataset," Sci. Rep., vol. 14, no. 1, p. 14429, 2024, doi: 10.1038/s41598-024-65276-x.

[11] S. A. Salloum, K. M. Alomari, A. M. Alfaisal, R. A. Aljanada, and A. Basiouni, "Emotion recognition for enhanced learning: using AI to detect students' emotions and adjust teaching methods," Smart Learn. Environ., vol. 12, no. 1, p. 21, 2025, doi: 10.1186/s40561-025-00374-5.

[12] T.-W. Kim and K.-C. Kwak, "Speech Emotion Recognition Using Deep Learning Transfer Models and Explainable Techniques," 2024. doi: 10.3390/app14041553.

[13] H. Alharbi, "Explainable feature selection and deep learning-based emotion recognition in virtual reality using eye tracker and physiological data," Front. Med., vol. 11, p. 1438720, 2024, doi: 10.3389/fmed.2024.1438720.

[14] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human Emotion Recognition: Review of Sensors and Methods.," Sensors (Basel)., vol. 20, no. 3, Jan. 2020, doi: 10.3390/s20030592.

[15] R. Guo, H. Guo, L. Wang, M. Chen, D. Yang, and B. Li, "Development and application of emotion recognition technology - a systematic literature review.," BMC Psychol., vol. 12, no. 1, p. 95, Feb. 2024, doi: 10.1186/s40359-024-01581-4.

[16] S. M. George and P. Muhamed Ilyas, "A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise," Neurocomputing, vol. 568, p. 127015, 2024, doi: https://doi.org/10.1016/j.neucom.2023.127015.

[17] A. Pentari, G. Kafentzis, and M. Tsiknakis, "Speech emotion recognition via graph-based representations," Sci. Rep., vol. 14, no. 1, p. 4484, 2024, doi: 10.1038/s41598-024-52989-2.

[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, p. 22, 2023, doi: 10.1186/s13636-023-00289-4.

[19] A. Chakhtouna, S. Sekkate, and A. Adib, "Efficient bimodal emotion recognition system based on speech/text embeddings and ensemble learning fusion," Ann. Telecommun., vol. 80, no. 5, pp. 379–399, 2025, doi: 10.1007/s12243-025-01088-y.

[20] C. Barhoumi and Y. BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation," Artif. Intell. Rev., vol. 58, no. 2, p. 49, 2024, doi: 10.1007/s10462-024-11065-x.

[21] M. Kaur and M. Kumar, "Facial emotion recognition: A comprehensive review," Expert Syst., vol. 41, no. 10, p. e13670, 2024.

[22] O. Kalyta, O. Barmak, P. Radiuk, and I. Krak, "Facial Emotion Recognition for Photo and Video Surveillance Based on Machine Learning and Visual Analytics," 2023. doi: 10.3390/app13179890.

[23] Z.-Y. Huang et al., "A study on computer vision for facial emotion recognition," Sci. Rep., vol. 13, no. 1, p. 8425, 2023, doi: 10.1038/s41598-023-35446-4.

[24] B. Fang, Y. Zhao, G. Han, and J. He, "Expression-Guided Deep Joint Learning for Facial Expression Recognition," Sensors (Basel)., vol. 23, no. 16, Aug. 2023, doi: 10.3390/s23167148.

[25] A. Talukder and S. Ghosh, "Facial Image expression recognition and prediction system," Sci. Rep., vol. 14, no. 1, p. 27760, 2024, doi: 10.1038/s41598-024-79146-z.

[26] P. Thakur, N. Kaur, N. Aggarwal, and S. Singh, "A Comprehensive Review of Unimodal and Multimodal Emotion Detection: Datasets, Approaches, and Limitations," Expert Syst., vol. 42, no. 9, p. e70103, 2025.

[27] S. Gutiérrez, J. Fernández-Novales, T. Garde-Cerdán, S. Marín-San Román, J. Tardaguila, and M. P. Diago, "Multi-sensor spectral fusion to model grape composition using deep learning," Inf. Fusion, vol. 99, p. 101865, 2023, doi: https://doi.org/10.1016/j.inffus.2023.101865.

[28] E. M. G. Younis, S. Mohsen, E. H. Houssein, and O. A. S. Ibrahim, "Machine learning for human emotion recognition: a comprehensive review," Neural Comput. Appl., vol. 36, no. 16, pp. 8901–8947, 2024, doi: 10.1007/s00521-024-09426-2.

[29] J.-H. Lee, J.-Y. Kim, and H.-G. Kim, "Emotion Recognition Using EEG Signals and Audiovisual Features with Contrastive Learning," Bioeng. (Basel, Switzerland), vol. 11, no. 10, Oct. 2024, doi: 10.3390/bioengineering11100997.

[30] S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar, "A novel facial emotion recognition model using segmentation VGG-19 architecture," Int. J. Inf. Technol., vol. 15, no. 4, pp. 1777–1787, 2023, doi: 10.1007/s41870-023-01184-z.

[31] R. Baazeem, "Explainable cross-domain emotion recognition using non-linear optimization and multimodal feature fusion based deep learning model," Int. J. Inf. Technol., 2025, doi: 10.1007/s41870-025-02930-1.

[32] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," Sci. Rep., vol. 14, no. 1, p. 13126, 2024.

[33] P. Koromilas and T. Giannakopoulos, "Deep Multimodal Emotion Recognition on Human Speech: A Review," 2021. doi: 10.3390/app11177962.

[34] B. Akande, "Multimodal Deep Learning for Emotion Detection from Text, Audio, and Facial Expressions," no. April, 2025.

[35] A.-L. Cîrneanu, D. Popescu, and D. Iordache, "New Trends in Emotion Recognition Using Image Analysis by Neural Networks, A Systematic Review," 2023. doi: 10.3390/s23167092.

[36] A. R. Khan, "Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges," 2022. doi: 10.3390/info13060268.

[37] S. Son and Y. Jeong, "Face and Voice Recognition-Based Emotion Analysis System (EAS) to Minimize Heterogeneity in the Metaverse," Appl. Sci., vol. 15, no. 2, 2025, doi: 10.3390/app15020845.

[38] B. T. Atmaja and A. Sasou, "Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations," 2022. doi: 10.3390/s22176369.

[39] Y. Zhao and X. Shu, "Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC)," Sci. Rep., vol. 13, no. 1, p. 20398, 2023.

[40] C. Xu, Y. Liu, W. Song, Z. Liang, and X. Chen, "A New Network Structure for Speech Emotion Recognition Research," Sensors (Basel)., vol. 24, no. 5, Feb. 2024, doi: 10.3390/s24051429.

[41] D. Resende Faria, A. I. Weinberg, and P. P. Ayrosa, "Multimodal Affective Communication Analysis: Fusing Speech Emotion and Text Sentiment Using Machine Learning," 2024. doi: 10.3390/app14156631.

[42] K. Kaur and P. Singh, "Trends in speech emotion recognition: a comprehensive survey," Multimed. Tools Appl., vol. 82, no. 19, pp. 29307–29351, 2023.

[43] 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 (Basel), vol. 25, no. 10, Oct. 2023, doi: 10.3390/e25101440.

[44] Y. Zeng, J.-W. Zhang, and J. Yang, "Multimodal emotion recognition in the metaverse era: New needs and transformation in mental health work," World J. Clin. cases, vol. 12, no. 34, p. 6674, 2024.

[45] A. A. Wafa, M. M. Eldefrawi, and M. S. Farhan, "Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning," J. Big Data, vol. 12, no. 1, p. 210, 2025.

[46] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, "MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion," Sci. Rep., vol. 15, no. 1, p. 5473, 2025.

[47] F. Bao, M. Neumann, and N. T. Vu, "CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2019-Septe, pp. 2828–2832, 2019, doi: 10.21437/Interspeech.2019-2293.

[48] N. Saleem et al., "DeepCNN: Spectro-temporal feature representation for speech emotion recognition," CAAI Trans. Intell. Technol., vol. 8, no. 2, pp. 401–417, 2023, doi: 10.1049/cit2.12233.

[49] L. Wijayasingha and J. A. Stankovic, "Robustness to noise for speech emotion classification using CNNs and attention mechanisms," Smart Heal., vol. 19, p. 100165, 2021, doi: https://doi.org/10.1016/j.smhl.2020.100165.

[50] M. F. Almufareh, S. Kausar, M. Humayun, and S. Tehsin, "A Conceptual Model for Inclusive Technology: Advancing Disability Inclusion through Artificial Intelligence," J. Disabil. Res., vol. 3, no. 1, pp. 1–11, 2024, doi: 10.57197/JDR-2023-0060.

[51] R. Malviya and S. Rajput, "AI-Driven Innovations in Assistive Technology for People with Disabilities BT - Advances and Insights into AI-Created Disability Supports," R. Malviya and S. Rajput, Eds., Singapore: Springer Nature Singapore, 2025, pp. 61–77. doi: 10.1007/978-981-96-6069-8_4.

[52] N. Simić et al., "Enhancing Emotion Recognition through Federated Learning: A Multimodal Approach with Convolutional Neural Networks," 2024. doi: 10.3390/app14041325.

[53] F. Makhmudov, A. Kultimuratov, and Y.-I. Cho, "Enhancing Multimodal Emotion Recognition through Attention Mechanisms in BERT and CNN Architectures," 2024. doi: 10.3390/app14104199.

[54] S. Garcia, F. Gomez-Donoso, and M. Cazorla, "Enhancing Human–Robot Interaction: Development of Multimodal Robotic Assistant for User Emotion Recognition," 2024. doi: 10.3390/app142411914.

[55] K. Chemnad, "Digital accessibility in the era of artificial intelligence — Bibliometric analysis and systematic review," Front. Artif. Intell., 2022.

[56] M. M. Terras, D. Jarrett, and S. A. McGregor, "The Importance of Accessible Information in Promoting the Inclusion of People with an Intellectual Disability," 2021. doi: 10.3390/disabilities1030011.

[57] H. M. Salman and R. T. Rasheed, "Smart Door for Handicapped People via Face Recognition and Voice Command Technique," Eng. Technol. J., vol. 39, no. 1B, pp. 222–230, 2021, doi: 10.30684/etj.v39i1b.1719.

[58] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical AI," Nat. Med., vol. 28, no. 9, pp. 1773–1784, 2022, doi: 10.1038/s41591-022-01981-2.

[59] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," Int. J. Appl. Earth Obs. Geoinf., vol. 112, p. 102926, 2022, doi: https://doi.org/10.1016/j.jag.2022.102926.

[60] R. Haque, N. Islam, M. Tasneem, and A. K. Das, "Multi-class sentiment classification on Bengali social media comments using machine learning," Int. J. Cogn. Comput. Eng., vol. 4, pp. 21–35, 2023, doi: https://doi.org/10.1016/j.ijcce.2023.01.001.

[61] W. Chai and G. Wang, "Deep Vision Multimodal Learning: Methodology, Benchmark, and Trend," 2022. doi: 10.3390/app12136588.

[62] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," Inf. Fusion, vol. 91, pp. 424–444, Mar. 2023, doi: 10.1016/j.inffus.2022.09.025.

[63] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision : advances, trends, applications, and datasets," Vis. Comput., vol. 38, no. 8, pp. 2939–2970, 2022, doi: 10.1007/s00371-021-02166-7.

[64] A. Mukherjee and M. Shrivastava, "Lost in Translation? Found in Evaluation: A Comprehensive Survey on Sentence-Level Translation Evaluation," ACM Comput. Surv., vol. 58, no. 1, Sep. 2025, doi: 10.1145/3735970.

[65] P. J. Vaz, J. M. F. Rodrigues, and P. J. S. Cardoso, "Affective Computing Emotional Body Gesture Recognition: Evolution and the Cream of the Crop," IEEE Access, vol. 13, no. September, pp. 192871–192890, 2025, doi: 10.1109/ACCESS.2025.3630563.

[66] S. Learn, "scikit-learn Machine Learning in Python," scikit-learn. Accessed: Jul. 02, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html?utm_source=chatgpt.com#

[67] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," Technometrics, vol. 64, no. 2, pp. 166–176, Apr. 2022, doi: 10.1080/00401706.2021.1921037.

[68] J. Park, E. Byon, Y. M. Ko, and S. Shashaani, "Strata design for variance reduction in stochastic simulation," Technometrics, vol. 67, no. 2, pp. 203–214, 2025.

[69] H. Liu and M. Cocea, "Semi-random partitioning of data into training and test sets in granular computing context," Granul. Comput., vol. 2, no. 4, pp. 357–386, 2017, doi: 10.1007/s41066-017-0049-2.

[70] O. Yazdanpanah, A. Formisano, M. Chang, and B. Mohebi, "Fragility curves for seismic damage assessment in regular and irregular MRFs using improved wavelet-based damage index," Measurement, vol. 182, p. 109558, 2021, doi: https://doi.org/10.1016/j.measurement.2021.109558.

[71] R. Cantini et al., "Block size estimation for data partitioning in HPC applications using machine learning techniques," J. Big Data, vol. 11, no. 1, p. 19, 2024, doi: 10.1186/s40537-023-00862-w.

[72] O. Elharrouss et al., "Task-based Loss Functions in Computer Vision: A Comprehensive Review," Apr. 01, 2025. doi: 10.48550/arXiv.2504.04242.

[73] J.-S. Hwang, S.-S. Lee, J.-W. Gil, and C.-K. Lee, "Determination of Optimal Batch Size of Deep Learning Models with Time Series Data," 2024. doi: 10.3390/su16145936.

[74] J. Liu, H. Wang, M. Sun, and Y. Wei, "Graph based emotion recognition with attention pooling for variable-length utterances," Neurocomputing, vol. 496, pp. 46–55, 2022, doi: https://doi.org/10.1016/j.neucom.2022.05.007.

[75] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light-weight deep neural ensemble model using handcrafted features," Sci. Rep., vol. 15, no. 1, p. 11824, 2025, doi: 10.1038/s41598-025-95734-z.

[76] Y. Ahn, S. Han, S. Lee, and J. W. Shin, "Speech Emotion Recognition Incorporating Relative Difficulty and Labeling Reliability," Sensors (Basel)., vol. 24, no. 13, Jun. 2024, doi: 10.3390/s24134111.

[77] W. J. C, G. J. W. Kathrine, V. Shanmuganathan, and S. Sumathi, "Improved optimizer with deep learning model for emotion detection and classification," vol. 21, no. June, pp. 6631–6657, 2024, doi: 10.3934/mbe.2024290.

[78] A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: a comprehensive review," Multimed. Tools Appl., vol. 84, no. 21, pp. 24079–24190, 2025, doi: 10.1007/s11042-024-20011-6.

[79] M. Abdullahi et al., "A systematic literature review of visual feature learning: deep learning techniques, applications, challenges and future directions," Multimed. Tools Appl., vol. 84, no. 19, pp. 20439–20496, 2025, doi: 10.1007/s11042-024-19823-3.

[80] P. Amalia et al., "Implementasi Algoritma Convolutional Neural Network (CNN) Dengan Optimizer Adam Dalam Deteksi Emosional Pada Wajah Manusia," J. VOKASI Tek., vol. 3, no. 1, pp. 13–22, 2025, [Online]. Available: https://mentech.id/jurnal/index.php/juvotek/article/view/60

[81] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning.," Sci. Rep., vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[82] J. Li, "Area under the ROC Curve has the most consistent evaluation for binary classification.," PLoS One, vol. 19, no. 12, p. e0316019, 2024, doi: 10.1371/journal.pone.0316019.

[83] P. Mishra, A. S. Verma, P. Chaudhary, and A. Dutta, "Emotion Recognition from Facial Expression Using Deep Learning Techniques," 2024 IEEE 9th Int. Conf. Converg. Technol. I2CT 2024, 2024, doi: 10.1109/I2CT61223.2024.10543313.

[84] D. Resha, P. Pamungkas, and B. H. Prasetio, "Implementasi Ekstraksi Gammatone-Frequency Cepstral Coefficient dan Klasifikasi Hidden Markov Model dalam Identifikasi Emosi Menggunakan Suara Jantung," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 9, no. 3, pp. 1–10, 2025, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/14533

[85] M. M. Taye, "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions," 2023. doi: 10.3390/computation11030052.

[86] A. Al Bataineh, D. Kaur, M. Al-khassaweneh, and E. Al-sharoa, "Automated CNN Architectural Design: A Simple and Efficient Methodology for Computer Vision Tasks," 2023. doi: 10.3390/math11051141.

[87] G. Ortega-Flores, G. Altamirano-Escobedo, D. Mercado-Ravell, and E. Bayro-Corrochano, "Quaternion CNN in Deep Learning Processing for EEG with Applications to Brain Disease Detection," 2025. doi: 10.3390/app152111526.

[88] F. A. Albalooshi and M. R. Qader, "Deep Learning Algorithm for Automatic Classification of Power Quality Disturbances," 2025. doi: 10.3390/app15031442.

[89] G. Hernández-Nava, S. Salazar-Colores, E. Cabal-Yepez, and J.-M. Ramos-Arreguín, "Parallel Ictal-Net, a Parallel CNN Architecture with Efficient Channel Attention for Seizure Detection," 2024. doi: 10.3390/s24030716.

[90] I. Rakhmatulin, M.-S. Dao, A. Nassibi, and D. Mandic, "Exploring Convolutional Neural Network Architectures for EEG Feature Extraction," 2024. doi: 10.3390/s24030877.

[91] R. Begazo, A. Aguilera, I. Dongo, and Y. Cardinale, "A Combined CNN Architecture for Speech Emotion Recognition," 2024. doi: 10.3390/s24175797.

[92] X. Liu, C. Cao, and S. Duan, "A Low-Power Hardware Architecture for Real-Time CNN Computing," 2023. doi: 10.3390/s23042045.

[93] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J. Big Data, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[94] I. D. Mienye and T. G. Swart, "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications," 2024. doi: 10.3390/info15120755.