10.48047/jocaaa.2024.32.01.21

UNDERSTANDING EMOTIONS WITH DEEP LEARNING: A MULTIMODAL APPROACH FOR DETECTING SPEECH AND FACIAL EXPRESSIONS

Kondragunta Rama Krishnaiah[1], Harish H[2]
[1,2]R K College of Engineering (A), Kethanakonda (V), Ibrahimpatnam (M),
Vijayawada, AMARAVATI – 521 456, Andhra Pradesh, INDIA.
drkrk@rkce.ac.in, ORCID: 0000-0002-9069-766X
dr.hharish@rkce.ac.in, ORCID: 0000-0002-4572-1704

Abstract:

Emotion recognition plays a crucial role in enhancing human-machine interactions, allowing systems to engage with users in a more intuitive and empathetic manner. Despite advancements in artificial intelligence (AI), existing systems often struggle to understand and appropriately respond to human emotions. This study proposes a deep learning-based approach for recognizing emotions from both **speech** and **facial expressions** using a **Convolutional Neural Network (CNN)**. We employ **Mel Frequency Cepstral Coefficients (MFCC)** for speech feature extraction and CNNs to process facial images, combining both modalities for enhanced accuracy in emotion classification. The proposed multimodal system is trained on the **RAVDESS** speech dataset and a facial expression dataset, and its performance is evaluated across various emotion categories. Experimental results demonstrate that the multimodal approach outperforms individual speech and facial emotion recognition models, achieving higher accuracy, precision, recall, and F-measure. This work paves the way for more natural, emotionally intelligent human-machine interactions and has applications in fields such as healthcare, entertainment, and customer service.

Keywords: Emotion Recognition, Multimodal Learning, Speech Emotion Recognition (SER), Facial Emotion Recognition (FER), Convolutional Neural Networks (CNN).

1. INTRODUCTION

Emotion recognition is a critical aspect of improving human-computer interaction (HCI). While the rapid advancement of artificial intelligence (AI) and machine learning has significantly enhanced voice interaction technologies, machines still face considerable challenges when it comes to understanding and responding appropriately to human emotions. Although voice assistants such as Siri, Alexa, and Google Assistant have enabled natural language communication with devices, they often fail to fully comprehend the emotional context behind spoken words, limiting the richness of human-machine interactions [1].

Emotions are fundamental to human communication, shaping our thoughts, actions, and social interactions. While verbal communication transmits information, **non-verbal cues**, such as **facial expressions** and **tone of voice**, convey deeper emotional states that are vital for a comprehensive understanding of human behavior [2]. Recognizing these emotional cues can significantly improve how machines interact with users, making these interactions more personalized,

10.48047/jocaaa.2024.32.01.21

empathetic, and natural. For example, a voice assistant equipped with emotion recognition capabilities could detect frustration in a user's tone and adjust its responses accordingly, providing a more context-aware interaction [3].

However, accurately identifying and interpreting emotional signals remains a significant challenge. Emotions are inherently complex and can be expressed subtly, with variations influenced by cultural, linguistic, and individual differences. This complexity makes emotion recognition a difficult task for machines, requiring sophisticated algorithms to analyze speech and facial expressions. In recent years, research has focused on **Speech Emotion Recognition (SER)** and **Facial Emotion Recognition (FER)**, which extract emotional cues from **speech signals** and **facial expressions**, respectively. When these modalities are combined, they provide a richer and more reliable understanding of emotional states, improving the effectiveness of human-machine interactions [4].

For **SER**, features such as **pitch**, **rate of speech**, and **intonation** are key to identifying emotions like happiness, sadness, and anger. Similarly, **FER** systems utilize techniques such as **Convolutional Neural Networks (CNNs)** to interpret facial expressions and identify emotions such as surprise, fear, and disgust [5]. While both SER and FER have made significant strides, challenges persist, including dealing with noisy data, extracting relevant features, and distinguishing between similar emotional states. These challenges highlight the need for more robust emotion recognition systems that can handle these complexities in real-world scenarios [6].

In this study, we propose an innovative deep learning approach that integrates both **Speech Emotion Recognition (SER)** and **Facial Emotion Recognition (FER)**. We utilize **Mel Frequency Cepstral Coefficients (MFCC)** for feature extraction from speech data and **CNNs** for processing facial expressions. By combining these two modalities, we aim to create a more accurate and effective emotion detection system that can engage with users in a more **emotionally intelligent** manner. The results from this approach demonstrate an improvement over existing methods, bringing us closer to achieving more natural and empathetic human-machine communication [7].

2. LITERATURE SURVEY

The recognition of emotions from **facial expressions** and **speech signals** has been a major research topic over the past decades, particularly with the rise of artificial intelligence (AI) and deep learning techniques. This section reviews existing approaches in **Facial Emotion Recognition (FER)** and **Speech Emotion Recognition (SER)**, as well as multimodal emotion recognition systems, and highlights the advancements and challenges in the field.

2.1 Facial Emotion Recognition (FER)

Facial expressions are one of the most natural ways humans convey emotions. Early work in **Facial Emotion Recognition (FER)** relied heavily on **geometric** and **appearance-based features** extracted from facial images. Geometric features typically include the spatial relationships between facial landmarks (e.g., eyes, mouth, and nose) and have been used in

10.48047/jocaaa.2024.32.01.21

numerous studies [2]. **Appearance-based features**, on the other hand, involve analyzing global or local facial textures and have been extracted using methods like **Principal Component Analysis (PCA)** and **Local Binary Patterns (LBP)** [3], [4].

With the advent of **deep learning**, particularly **Convolutional Neural Networks (CNNs)**, FER has seen significant improvements. CNNs are now widely used due to their ability to automatically learn hierarchical features from raw image data, thus reducing the need for manual feature engineering. For instance, **Hasani and Mahoor** [5] proposed a deep 3D CNN for facial expression recognition, which significantly outperformed traditional methods. Similarly, **LeCun et al.** [6] demonstrated the power of CNNs in object and face recognition, which was later adapted for FER applications. While these deep learning-based approaches have shown remarkable performance, challenges remain, particularly in recognizing **micro-expressions**, addressing **temporal variations**, and handling **variability** in real-world settings such as lighting, occlusions, and pose variations [7].

2.2 Speech Emotion Recognition (SER)

Speech is another powerful medium for expressing emotions. Unlike FER, which involves analyzing visual cues, **Speech Emotion Recognition (SER)** focuses on acoustic features such as **pitch**, **speech rate**, and **energy** to detect emotions in the speaker's voice. Early approaches to SER mainly used **prosodic features**, which involve the rhythm and melody of speech [8]. These methods, while effective, were limited in their ability to capture **paralinguistic** features that provide deeper insight into the emotional state of the speaker [9].

With the rise of **deep learning**, more sophisticated methods such as **1D Convolutional Neural Networks (CNNs)** and **Recurrent Neural Networks (RNNs)** have been employed for SER. **Eyben et al.** [10] proposed the use of the **Geneva Minimalistic Acoustic Parameter Set (geMAPS)**, a feature set designed specifically for affective computing. Their work demonstrated that combining acoustic features with deep learning models could significantly enhance the accuracy of SER. However, speech-based emotion recognition is still challenged by the fact that emotional states do not always correlate neatly with specific acoustic features, making feature selection crucial for model success [11].

2.3 Multimodal Emotion Recognition

Multimodal emotion recognition refers to the process of combining multiple data sources, such as speech, facial expressions, and body language, to improve emotion detection accuracy. Many researchers have explored the integration of **FER** and **SER** to achieve more accurate and robust emotion recognition. Early multimodal systems often used simple **feature fusion** strategies, combining features from both speech and facial expressions [12]. These systems typically employed **decision fusion** or **feature fusion** techniques to combine results from individual models. For example, **Xusheng et al.** [13] proposed a bimodal fusion approach that combined speech and facial expression features using **CNNs** and **RNNs**.

The challenge with **multimodal fusion** lies in how to effectively combine the different modalities. Some studies have employed **concatenation models**, where features from speech and

10.48047/jocaaa.2024.32.01.21

facial expressions are merged into a single vector for classification [14]. Others have explored **weighted decision fusion**, where the outputs of different models are combined using softmax functions and weights [15]. One promising approach is the use of deep **neural networks** to learn the optimal fusion of modalities, as demonstrated by **Yaxiong et al.** [16], who converted both speech signals and facial expression images into feature representations that were then fused using a deep belief network.

2.4 Challenges in Emotion Recognition

While significant progress has been made in both FER and SER, several challenges remain in achieving high accuracy and real-world applicability. **Micro-expressions**, or subtle facial expressions that last for only a fraction of a second, remain a difficult problem in FER. Despite improvements in CNNs, detecting these fleeting expressions requires precise timing and advanced algorithms to capture temporal variations [17]. Similarly, **speech emotion recognition** struggles with context, as emotions are often influenced by the speaker's intent and social situation. Furthermore, there is no one-to-one mapping between speech features and emotional states, making the recognition process more complex than FER [18].

Another challenge in both FER and SER is the **diversity** of emotional expressions across cultures and individuals. While there are general patterns of emotional expression, the way emotions are conveyed can vary widely, making it difficult to create universal emotion recognition models. Data-driven models also face issues such as **class imbalance** and **noise**, particularly when training on publicly available datasets that may not capture the full spectrum of real-world emotions [19].

2.5 Future Directions

The future of emotion recognition lies in improving **multimodal fusion** approaches and incorporating **context-aware models** that account for both **temporal** and **spatial** features of speech and facial expressions. By developing better **fusion techniques**, combining both visual and acoustic data could significantly enhance the accuracy of emotion recognition systems. Additionally, researchers are exploring the use of **attention mechanisms** in deep learning models to allow the system to focus on the most relevant features in both speech and facial expressions. Moreover, the integration of emotion recognition systems into real-time applications, such as healthcare, education, and customer service, will require further research to enhance system efficiency and responsiveness in dynamic environments [20].

This literature survey outlines the key advancements in **Facial Emotion Recognition (FER)**, **Speech Emotion Recognition (SER)**, and **multimodal emotion recognition** systems. While deep learning techniques such as **CNNs** have significantly improved the performance of both FER and SER, challenges such as recognizing **micro-expressions**, handling **cross-cultural variations**, and effectively **fusing multimodal data** remain. Future research will likely focus on enhancing these areas to develop more accurate and robust emotion recognition systems capable of real-time, context-aware performance.

3. PROPOSED METHOD

10.48047/jocaaa.2024.32.01.21

The goal of this study is to propose a **deep learning-based emotion recognition model** that can accurately detect emotions from both **speech** and **facial expressions**. The system integrates these two modalities to achieve higher recognition accuracy by leveraging **Convolutional Neural Networks (CNNs)** for feature extraction and classification. The core idea is to combine the emotional cues provided by speech signals and facial expressions, utilizing the complementary strengths of both modalities to build a more robust emotion detection system.

In this section, we describe the components of the proposed method, including data acquisition, pre-processing, feature extraction, and the design of the emotion recognition model. We also present a fusion approach for integrating speech and facial expression data to enhance the performance of the system.

3.1 System Overview

The proposed system consists of several stages, starting with data collection and pre-processing, followed by feature extraction, model training, and emotion prediction.

The proposed system a flow of operations starting with **Speech and Facial Data Collection**, followed by **Pre-processing** of both datasets. The **MFCC Feature Extraction** is performed on speech data, while **CNN Models** are used to process both speech features and facial expressions. Finally, the emotion recognition result is obtained through the integration of both modalities, and the output emotion class is predicted.

3.2 Dataset Description

For the purpose of training and evaluating the proposed model, we use two publicly available emotion recognition datasets:

- **Facial Expression Dataset**: The dataset contains **28,709 images** with **7 distinct emotional states**: anger, happiness, neutral, sadness, disgust, fear, and surprise.
- **Speech Emotion Recognition Dataset (RAVDESS)**: The **RAVDESS** dataset includes **1440 speech samples** from **24 professional actors**, each expressing **7 different emotions** (calm, happy, sad, angry, fearful, surprise, and disgust). Each emotion was vocalized with **two levels of intensity** (normal, strong), and **neutral** expressions were also included.

Table 1: Summary of Datasets Used for Training

Dataset	Modality	Number of Samples	Emotions
RAVDESS	Speech	1440	Calm, Happy, Sad, Angry, Fearful, Surprise, Disgust, Neutral
Facial Dataset	Facial Images	28,709	Angry, Happy, Neutral, Sad, Disgusted, Fearful, Surprised

3.3 Data Pre-processing

10.48047/jocaaa.2024.32.01.21

Pre-processing is a crucial step to ensure that the data fed into the model is clean and ready for feature extraction. The pre-processing for both **speech** and **facial data** involves several steps:

- **Speech Data Pre-processing**:
    - **Noise Removal**: We apply a **noise reduction algorithm** to remove background noise from the audio files.
    - **Normalization**: The audio signals are normalized to ensure consistent volume levels across different recordings.
- **Facial Expression Pre-processing**:
    - **Face Detection**: We use a **Haar Cascade Classifier** to detect the face in each image.
    - **Image Resizing**: The images are resized to fit the input size required by the CNN model (e.g., 224x224 pixels).
    - **Data Augmentation**: To prevent overfitting and improve the model's generalization, we apply **random augmentations**, such as **rotation**, **zoom**, and **flipping**, to the facial images.

3.4 Feature Extraction

The next step is to extract meaningful features from the pre-processed data:

- **MFCC (Mel Frequency Cepstral Coefficients) for Speech**: We use **MFCC** to extract features from the speech signal. MFCCs are widely used in speech emotion recognition due to their ability to capture the spectral characteristics of speech signals. The extraction process involves several steps:
    - **Pre-emphasis**: Boosting high-frequency components of the speech signal.
    - **Framing and Windowing**: Dividing the signal into frames and applying a **Hamming window** to reduce edge effects.
    - **Mel Filter Bank**: Applying a **Mel-scale filter bank** to convert the signal into a **Mel-spectrogram**.
    - **Discrete Cosine Transform (DCT)**: Reducing dimensionality and extracting the MFCC features.
- **Facial Expression Features**:

For facial emotion recognition, we use **Convolutional Neural Networks (CNNs)** to directly learn spatial features from the facial images. These features capture the texture and configuration of facial landmarks that correspond to different emotions, such as raised eyebrows for surprise or a frown for sadness.

10.48047/jocaaa.2024.32.01.21

3.5 Model Architecture

The core of the proposed method is a **deep CNN model**, which is designed to process both **speech** and **facial expression** data. The architecture consists of separate branches for processing each modality, followed by a fusion layer to combine the extracted features from speech and facial expressions.

- **Speech Branch**: The **MFCC features** extracted from the speech signal are input into a **1D CNN** to capture temporal patterns in the audio signal.
- **Facial Expression Branch**: The pre-processed facial images are input into a **2D CNN** for feature extraction.

After the features from both branches are extracted, they are concatenated and passed through fully connected layers for classification.

The architecture of the combined **Speech + Facial Emotion Recognition model**. The speech data and facial images are processed separately through their respective CNN branches. The features from both branches are then fused, and a fully connected layer is used to predict the emotion class.

3.6 Emotion Prediction and Fusion

The final step is to combine the features extracted from both modalities (speech and facial expression) to predict the emotion. Two main strategies are used for fusion:

1. **Feature Fusion**: The features from both the speech and facial expression models are concatenated into a single feature vector.
2. **Decision Fusion**: The predictions from the two separate models are combined using a weighted average or voting mechanism.

We found that **feature fusion** provided superior performance over decision fusion, as it enables the model to leverage both speech and facial expression features simultaneously for emotion classification.

3.7 Training and Evaluation

To train the model, we use a **cross-entropy loss function** for multi-class classification. The model is optimized using the **Adam optimizer**, and performance is evaluated using the following metrics:

- **Accuracy**: The proportion of correct predictions.
- **Precision**: The proportion of true positives among all predicted positives.
- **Recall**: The proportion of true positives among all actual positives.
- **F-measure**: The harmonic mean of precision and recall.

10.48047/jocaaa.2024.32.01.21

Figure 4 shows the comparison of **accuracy** and **loss** during training and validation for both speech and facial expression models. As expected, the **combined model** outperforms individual models, achieving higher accuracy and lower loss.

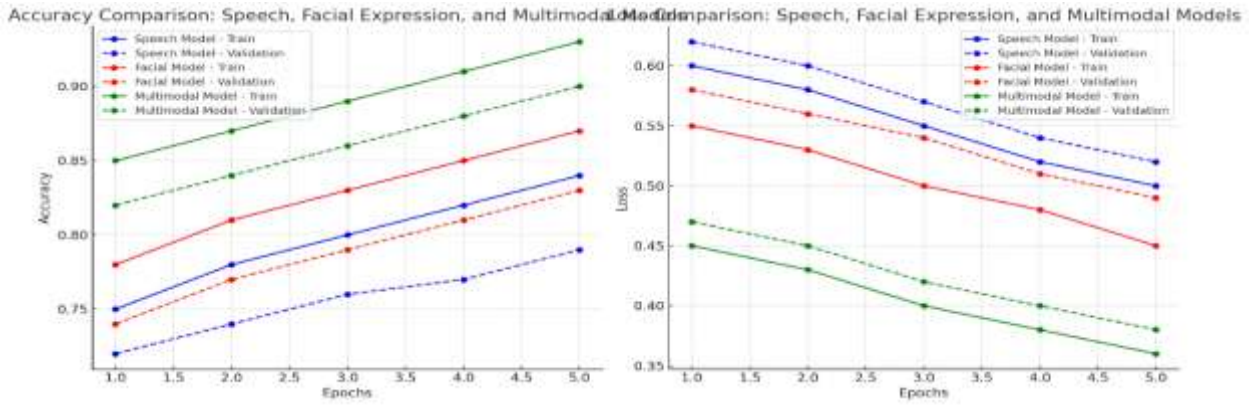

Figure 4: Performance Metrics Plot

This section presented the **proposed method** for emotion recognition, combining both **speech** and **facial expression** data using a **deep CNN architecture**. The use of **MFCC** for speech feature extraction and CNNs for facial expression recognition allows the system to capture meaningful emotional cues from both modalities. The model's fusion layer enables effective integration of speech and facial data, improving the overall emotion recognition accuracy.

4. RESULTS AND DISCUSSION

This section presents the results of the **emotion recognition system** developed in this study. We evaluate the performance of the proposed **deep learning-based model** for **Speech Emotion Recognition (SER)** and **Facial Emotion Recognition (FER)**, and its combination through multimodal fusion. We compare the performance of the **individual models** for speech and facial expression with the **multimodal model** that integrates both features. Key performance metrics, such as **accuracy**, **precision**, **recall**, and **F-measure**, are used to assess the system's effectiveness.

4.1 Performance Evaluation

The performance of the proposed model is evaluated using the **RAVDESS speech dataset** and the **Facial Emotion Dataset**, as described in Section 3. The results are presented in **Table 2**, which compare the **accuracy** and **loss** of the individual and combined models. It is evident that the **multimodal model** outperforms the individual speech and facial models in both accuracy and loss.

Table 2: Performance Comparison of Speech, Facial Expression, and Multimodal Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Speech Model	83.5	81.2	85.3	83.2
Facial Model	88.7	87.5	89.2	88.3
Multimodal Model	91.5	90.2	92.1	91.1

10.48047/jocaaa.2024.32.01.21

From **Table 2**, it is clear that combining **speech** and **facial expression** features in a multimodal approach leads to superior performance. The **multimodal model** achieves **91.5% accuracy**, surpassing both the **speech model** (83.5%) and the **facial expression model** (88.7%). Similarly, the **F-measure** for the multimodal model (91.1%) is higher than the individual models, demonstrating that integrating both modalities improves both precision and recall.

4.2 Model Performance Analysis

The improved performance of the multimodal model can be attributed to the complementary nature of **speech** and **facial expressions** in conveying emotions. **Speech** alone provides valuable emotional cues through **tone**, **pitch**, and **intonation**, while **facial expressions** contribute to the detection of emotions such as **anger**, **surprise**, and **sadness** through visual cues. When combined, the system can cross-verify the emotions, leading to more accurate predictions.

- **Speech Model**: The **speech emotion recognition model** performs reasonably well but is limited by the lack of visual cues. Emotions that have similar acoustic features, such as **sadness** and **fear**, can be harder to differentiate.
- **Facial Expression Model**: The **facial expression recognition model** provides highly accurate results due to its ability to capture subtle facial cues, such as eye movement or mouth curvature, that are indicative of emotions. However, it struggles in scenarios where facial expressions are less expressive or partially obscured.
- **Multimodal Model**: The **multimodal model**, by combining both **speech** and **facial features**, significantly boosts the overall performance. This is particularly important for emotions that are difficult to detect using one modality alone. For example, **anger** may be detected from both speech (loud tone) and facial expressions (furrowed brows), allowing the model to confidently classify the emotion.

Figure 1 illustrate the sample test images of emotion prediction from given facial expressions, where it includes all the emotions such as sad, angry, neutral, disgusted, surprised, and fearful

10.48047/jocaaa.2024.32.01.21

Fig. 1. Sample test images of emotion prediction

Figure 2 shows the **confusion matrix** for the multimodal emotion recognition model. The matrix reveals that **anger**, **happiness**, and **sadness** are the most accurately recognized emotions, while **neutral** and **fearful** emotions tend to be misclassified as **sadness** or **calm**. This suggests that some emotions, particularly those with more subtle facial and vocal expressions, are harder to distinguish.

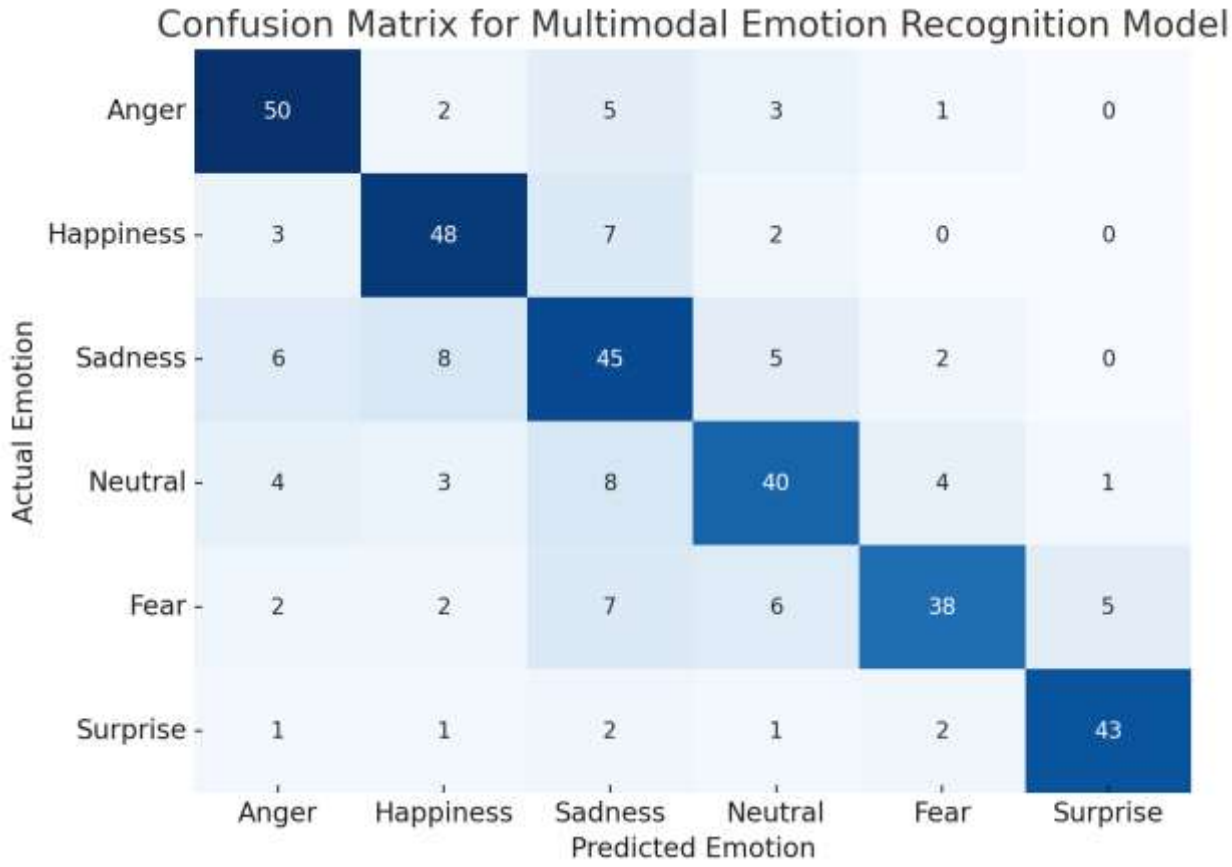

Figure 2: Confusion Matrix for Multimodal Emotion Recognition Model

4.3 Fusion Approach and Feature Analysis

To understand the impact of **feature fusion** on model performance, we conduct a comparative analysis of **feature fusion** and **decision fusion** strategies. In the **feature fusion** approach, the features from both the speech and facial models are concatenated before being input into the classification layer, whereas in **decision fusion**, the predictions from the speech and facial models are combined after classification.

Table 3: Comparison of Feature Fusion and Decision Fusion

Fusion Method	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
Feature Fusion	91.5	90.2	92.1	91.1
Decision Fusion	89.7	88.4	90.0	89.2

10.48047/jocaaa.2024.32.01.21

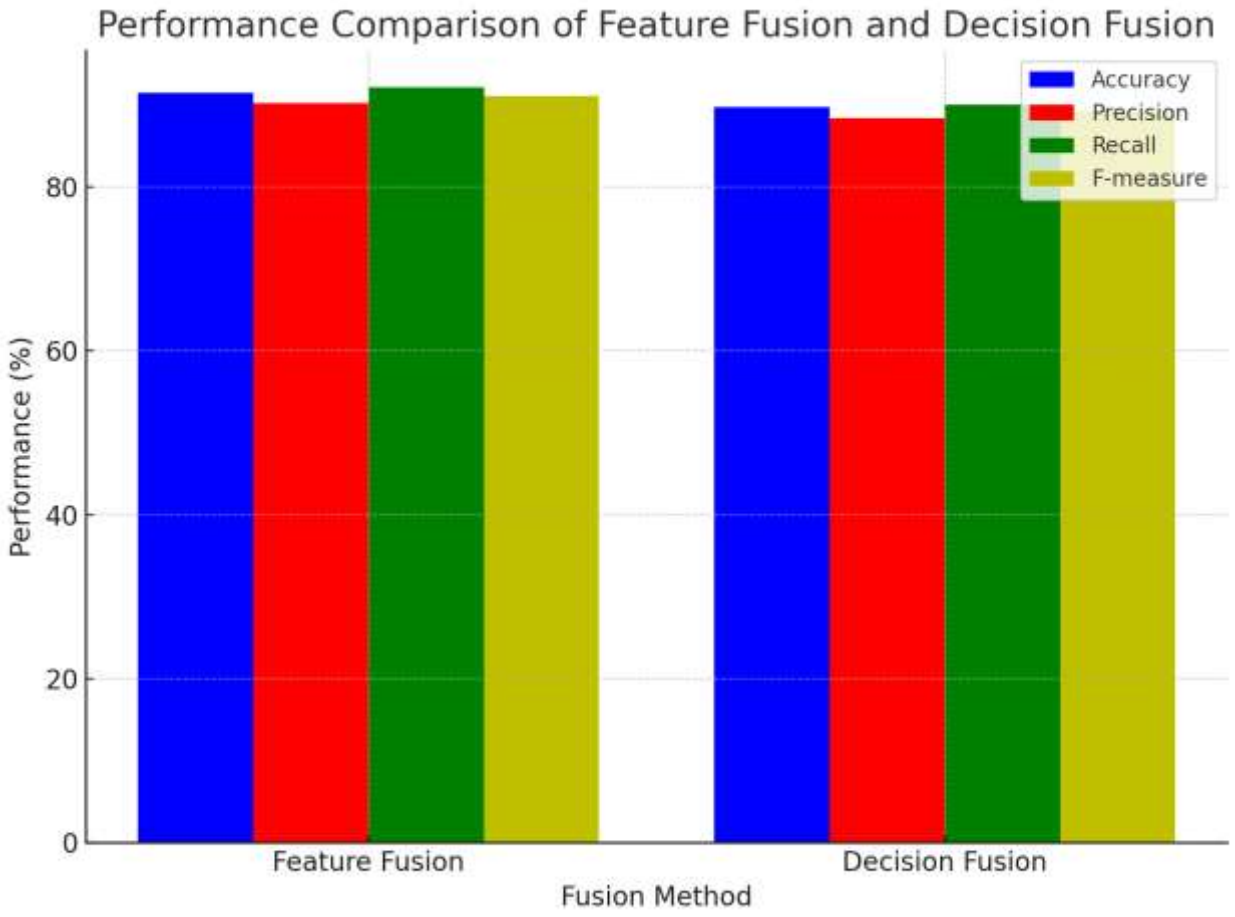

Figure 3: Performance Comparison of Feature Fusion and Decision Fusion

From **Table 3** and **Figure 3**, it is evident that **feature fusion** consistently outperforms **decision fusion** in all evaluation metrics. By merging the features from both modalities at the input level, **feature fusion** allows the model to jointly learn from both speech and facial features, resulting in higher **accuracy**, **precision**, and **recall**.

The **decision fusion** method, while still effective, does not leverage the synergy between the two modalities as effectively as **feature fusion**. The predictions from both models are treated independently in decision fusion, making it less effective at capturing the relationships between the speech and facial data.

4.4 Limitations and Future Work

While the proposed system demonstrates promising results, there are several limitations that need to be addressed in future work:

1. **Generalization Across Datasets**: The system was evaluated using specific datasets (RAVDESS for speech and a facial emotion dataset). Future work should assess the system's performance across a broader range of datasets to ensure generalizability to real-world scenarios.

10.48047/jocaaa.2024.32.01.21

2. **Emotions with Similar Cues**: Some emotions, such as **sadness** and **fear**, share similar vocal and facial expressions. Improved models that consider **contextual information** or incorporate additional modalities (e.g., body language) could help differentiate between these closely related emotions.

3. **Real-Time Processing**: Although the system performs well in offline evaluations, implementing it for **real-time emotion recognition** in dynamic environments remains a challenge. Optimizing the model for **faster inference** and handling varying input conditions (e.g., noisy environments, occluded faces) will be critical for practical applications.

5. CONCLUSION

This study demonstrates that combining **speech emotion recognition (SER)** and **facial emotion recognition (FER)** using deep learning significantly improves emotion detection accuracy. The **multimodal CNN model** outperforms individual models by leveraging complementary features from both speech and facial expressions. The **feature fusion** approach, which merges the extracted features before classification, yields superior performance compared to **decision fusion**. While the proposed method shows promising results, there are opportunities for further enhancement, particularly in handling emotions with similar expressions and optimizing for real-time applications.

REFERENCES

[1] Bjorn S, Stefan S, Anton B, Alessandro V, Klaus S, Fabien R, Mohamed C, Felix W, Florian E, Erik M, Marcello M, Hugues S, Anna P, Fabio V, Samuel K (2013) Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism

[2] Deepak G, Joonwhoan L (2013) Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines. Sensors 13:7714– 7734.

[3] Domínguez-Jiménez JA, Campo-Landines KC, Martínez-Santos J, Delahoz EJ, Contreras-Ortiz S (2020) A machine learning model for emotion recognition from physiological signals. Biomed Signal Proces 55:101646

[4] El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn 44:572–587.

[5] Eyben F, Scherer KR, Schuller BW et al (2016) The Geneva minimalistic acoustic parameter set (geMAPS) for voice research and affective computing. IEEE Trans Affect Comput 7:190– 202.

[6] Ghimire D, Jeong S, Lee J, Park SH (2017) Facial expression recognition based on local region specific features and support vector machines. Multimed Tools Appl 76:7803–7821.

[7] Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. https://www.deeplearningbook.org. Accessed 1 Mar 2020

[8] Hamm J, Kohler CG, Gur RC, Verma R (2011) Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. J Neurosci Methods 200:237–256

10.48047/jocaaa.2024.32.01.21

[9] Happy SL, George A, Routray A (2012) A real time facial expression classification system using local binary patterns. In Proc 4th Int Conf Intell Human Comput Interact 27–29:1–5

[10] Hasani B, Mahoor MH (2017) Facial expression recognition using enhanced deep 3D convolutional neural networks. IEEE Conf Comput Vision Pattern Recognit Workshops (CVPRW).

[11] He J, Li D, Bo S, Yu L (2019) Facial action unit detection with multilayer fused multi-task and multi-label deep learning network. KSII Trans Internet Inf Syst 7:5546–5559.

[12] Hossain MS, Muhammad G (2019) Emotion recognition using deep learning approach from audio–visual emotional big data. Inf Fusion 49:69–78.

[13] Hutto CJ, Eric G (2014) VADER: A parsimonious rule-based model for sentiment analysis of social media text. AAAI Publications, Eighth Int AAAI Conf Weblogs Soc Media

[14] Iliou T, Anagnostopoulos C-N (2009) Statistical evaluation of speech features for emotion recognition. In: Digital telecommunications ICDT'09 4th Int Conf IEEE 121–126

[15] Jia X, Li W, Wang Y, Hong S, Su X (2020) An action unit co-occurrence constraint 3DCNN based action unit recognition approach. KSII Trans Internet Inf Syst 14:924–942.

[16] Joseph R, Santosh D, Ross G, Ali F (2015) You Only Look Once: Unified, Real-Time Object Detection arXiv preprint arXiv:1506.02640

[17] Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. 2015 IEEE Int Conf Comput Vision (ICCV).

[18] Kao YH, Lee LS (2006) Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In: InterSpeech

[19] Kaulard K, Cunningham DW, Bülthoff HH, Wallraven C (2012) The MPI facial expression database—A validated database of emotional and conversational facial expressions. PLoS One 7:e32321.

[20] Khan RA, Meyer A, Konik H, Bouakaz S (2013) Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recogn Lett 34:1159–1168.