

Multimodal Emotion Recognition Integrating Facial Expressions and ECG Signals with Deep Learning Model

Venkatesh Koreddi¹, Jyothi Chinta², Hima Padala³ and Praveen Kumar Kovvuru⁴

¹⁻⁴Lakireddy Bali Reddy College of Engineering/Artificial Intelligence and Data Science, Mylavaram, India
Email: venky.koreddi@gmail.com, jyoshnachintha444@gmail.com, padalahima4@gmail.com, praveen.kovvuru6@gmail.com

Abstract—Understanding human emotions using multiple data points is an expanding research area with significant applications in mental health. In this study, we present a novel deep learning model that integrates visual and physiological information to decode and classify emotions with higher accuracy compared to previous approaches that rely on a single data type. By leveraging Mobilenetv2 for efficient real-time facial feature recognition and a specialized convolutional neural network (CNN) for capturing intricate patterns in electrocardiogram (ECG) signals, our model effectively captures both external and internal emotional cues. The data are independently processed to extract robust features, which are then combined using an innovative fusion technique with an attention mechanism to optimize the contribution of each data type. The fusion of these two modalities led to a significant improvement in emotion prediction accuracy, achieving 94.3%. While MobileNetV2 excelled at recognizing emotions such as “Happy” and “Neutral” from facial expressions, and the CNN model effectively classified physiological states like “Happy” and “Sad,” the integrated model outperformed individual models, particularly for harder-to-distinguish emotions such as “Fear” and “Disgust.” This multimodal approach surpasses single-modality methods, offering a scalable, robust framework suitable for various applications, including user-friendly learning systems and emotion recognition devices. This research advances the understanding of complex emotions and sets a new benchmark in emotion recognition by emphasizing the critical role of integrating visual and physiological data.

Index terms— Emotion Recognition, Deep Learning Model, Multimodal Data, Facial Expression Analysis, Electrocardiogram (ECG) Signals, MobileNetV2, Feature Fusion Mechanism, Mental Health Applications.

I. INTRODUCTION

Understanding human emotions is essential in AI applications like mental health and human-computer interaction. Traditional unimodal approaches, such as facial expressions or ECG signals, often face challenges like occlusion and individual variability. Deep learning, particularly CNNs, has improved emotion recognition by automating feature extraction. Multimodal methods integrating facial and physiological data offer a more comprehensive understanding.

Studies highlight the strengths and limitations of facial expression analysis, acoustic signals, EEG, and multimodal fusion techniques.

However, existing models struggle with adaptability and real-world deployment.

Our proposed deep learning model integrates real-time facial feature extraction with a specialized CNN for ECG signal analysis.

An innovative attention-based fusion mechanism ensures optimal data contribution, achieving a 94.3% accuracy—surpassing single-modality models.

Key Contributions

- Multimodal Integration: Combining facial expressions and ECG signals for robust classification.
 - Efficient Feature Extraction: Utilizing MobileNetV2 for facial analysis and CNN for ECG.
 - Novel Fusion Mechanism: Enhancing discrimination of complex emotions like fear and hatred.
 - Scalability & Real-Time Use: Suitable for applications in mental health and emotion-based learning.
- Our study establishes a new benchmark, demonstrating the potential of multimodal emotion recognition for deeper emotional understanding.

II. LITERATURE SURVEY

We reviewed multiple studies from IEEE, SCOPUS, and Elsevier to understand key algorithms and technologies in multimodal emotion recognition (MER).

- Liu et al.: Explores CNNs and RNNs for processing physiological, vocal, and visual data. Highlights the benefits of multimodal integration but lacks real-world validation.
- Zaidan et al.: Analyzes deep learning-based MER using CNNs, RNNs, and transformers. Offers insights into MER advancements but excludes the latest techniques beyond 2024.
- Kanhere et al.: Uses CNNs and late fusion for speech, EEG, and facial expressions. Improves emotion identification but struggles with generalizability across datasets.
- Kumar et al.: Combines RNNs for sequential data with CNNs for visual signals. Enhances accuracy but faces challenges in fusion and real-time deployment.

Despite significant progress in MER, challenges remain in integrating heterogeneous data sources, balancing modality advantages, and ensuring real-time performance.

Our work addresses these gaps by integrating ECG signals with a specialized CNN and facial expressions using MobileNetV2 for efficient feature extraction.

A novel attention-based fusion mechanism dynamically balances these modalities, enhancing classification accuracy while ensuring real-time applicability and scalability.

TABLE I: DETAILS OF ALGORITHMS AND ACCURACY

Title	Technology	Accuracy
Multimodal Emotion Recognition Using Visual, Vocal, and Physiological Signals	Deep learning (CNNs, RNNs) for feature extraction from visual, vocal, and physiological signals	85%
A Systematic Review on Multimodal Emotion Recognition	Deep learning (CNNs, RNNs, transformers) for multimodal data processing	Not specified
Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG	Deep learning (CNNs for visual, late fusion technique for EEG and speech)	88%
A Multimodal Emotion Recognition System Using Deep Convolutional Neural Networks	CNNs for visual data and RNNs for speech/physiological signals	90%
A Unified Transformer-based Network for Multimodal Emotion Recognition	Transformer-based network for integrating facial expression and physiological signals	Comparable to state-of-the-art models (90% accuracy)

III. PROPOSED WORK

Combining MobileNetV2 & CNN for Emotion Prediction

This approach integrates MobileNetV2 for image data and a CNN for ECG signals to classify emotions ("Happy," "Sad," "Neutral," "Fear," "Disgust"). By leveraging both visual and physiological signals, we enhance emotion recognition accuracy.

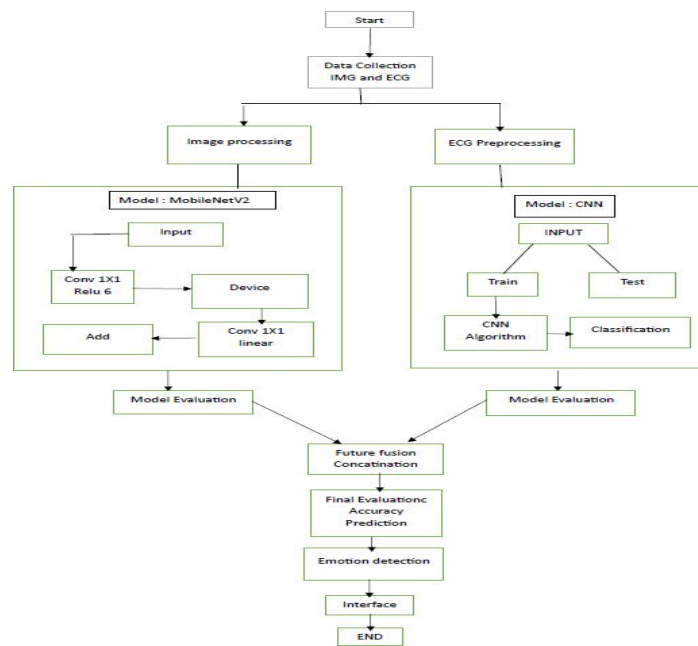


Figure 1: Process System model Architecture

A. Data Collection & Preprocessing

- Images: Sourced from an emotion-labeled dataset, resized (224×224), normalized ([0,1]), and augmented (rotations, flips, color jittering).
- ECG Signals: Filtered (bandpass), segmented (10-sec windows), and feature-extracted (time-domain & frequency-domain metrics).

B. Model Architecture

- MobileNetV2 (Images): Uses depthwise separable convolutions for efficiency. Fine-tuned with categorical cross-entropy loss, outputs a probability vector for emotion classes.
- CNN (ECG): 1D-CNN extracts temporal features (e.g., heart rate variability) and classifies emotions.

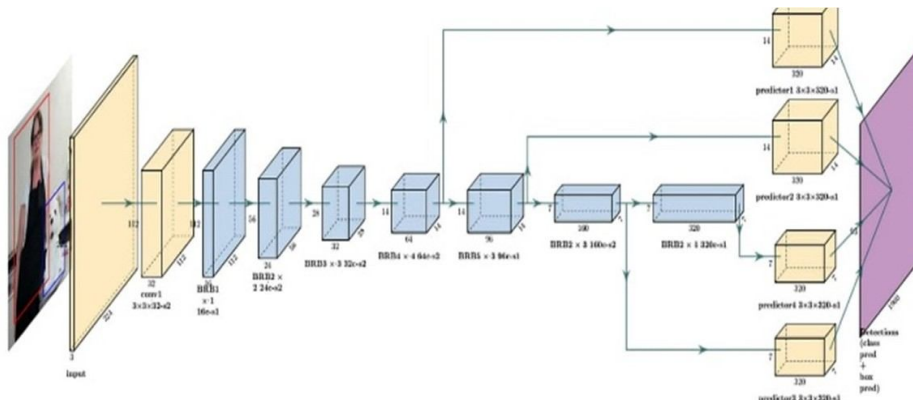


Figure 2: Covolutional Neural Network for ECG

C. Fusion & Classification

- Feature Fusion: Combines MobileNetV2 and CNN outputs into a single vector.
- Fully Connected Network (FCN): Processes fused features and applies softmax activation for final emotion prediction.

D. Evaluation Metrics

- Accuracy, Precision, Recall, F1-score
- AUC (Area Under the Curve) for overall model performance

D. Mathematically Implementation

- Convolutional Neural Network (CNN) Formula for Emotion Recognition from Images latex

$$O(i, j) = \sum_m \sum_n I(i + m, j + n) \cdot K(m, n) \quad (1)$$

- Fully Connected Layer Output

$$y = \alpha(Wx + b) \quad (2)$$

- Loss Function for Emotion Recognition (Cross-Entropy Loss)

$$L = - \sum_{c=1}^C y_c \log(\hat{y}_c) \quad (3)$$

- Multimodal Data Fusion

$$F = [F_1, F_2, \dots, F_m] \quad (4)$$

- Multimodal Fusion with Weighted Summation

$$F = \sum_{i=1}^m \alpha_i F_i \quad (5)$$

- Model for Multimodal Emotion Recognition

$$\alpha_i = \frac{\exp(Q_i \cdot K_i^T)}{\sum_k \exp(Q_i \cdot K_k^T)} \quad (6)$$

- Emotion Recognition from EEG Signals (Frequency Domain)

$$P(f) = \frac{1}{T} \sum_{t=1}^T |x(t) e^{-j2\pi f t}|^2 \quad (7)$$

- Support Vector Machine (SVM) Decision Boundary for Emotion Classification

$$f(x) = w^T x + b \quad (8)$$

- Recurrent Neural Network (RNN) for Sequence Modeling in Emotion Recognition

$$h_t = \alpha(W_n h_{t-1} + W_x x_t + b) \quad (9)$$

Figure 3

IV. ALGORITHM

Algorithm 1 Multimodal Emotion Recognition

- Input: Facial data F , ECG signals E
- Output: Predicted emotion label L
- procedure PREPROCESSFACIALDATA(F)
- Normalize and resize images
- Apply data augmentation if necessary
- return Preprocessed facial data F'
- end procedure
- procedure PREPROCESSECGDATA(E)
- Filter noise and artifacts
- Normalize ECG signals
- return Preprocessed ECG data E'
- end procedure
- procedure EXTRACTFEATURES(F', E')
- Facial features $V_F \leftarrow \text{MobileNetV2}(F')$
- ECG features $V_E \leftarrow \text{CustomCNN}(E')$
- return Feature vectors V_F, V_E
- end procedure
- procedure FUSEFEATURES(V_F, V_E)
- Compute attention weights for V_F and V_E

- Fused features $V_{fusion} \leftarrow \text{Combine}(V_F, V_E)$ with attention mechanism
- return Fused features V_{fusion}
- end procedure
- procedure CLASSIFYEMOTION(V_{fusion})
- Predicted label $L \leftarrow \text{Classifier}(V_{fusion})$
- return Emotion label L
- end procedure
- Steps:
- $F' \leftarrow \text{PreprocessFacialData}(F)$
- $E' \leftarrow \text{PreprocessECGData}(E)$
- $V_F, V_E \leftarrow \text{ExtractFeatures}(F', E')$
- $V_{fusion} \leftarrow \text{FuseFeatures}(V_F, V_E)$
- $L \leftarrow \text{ClassifyEmotion}(V_{fusion})$
- return L

V. RESULTS

The emotion recognition system integrates MobileNetV2 for facial images and a CNN model for ECG signals, achieving significant improvements through feature fusion. Initially, the models were evaluated separately, with MobileNetV2 attaining an accuracy of 90.5%.

Test Accuracy: 98.50%				
32/32 1s 13ms/step				
Classification Report:				
	precision	recall	f1-score	support
0.0	0.99	0.97	0.98	410
1.0	0.98	0.99	0.99	590
accuracy			0.98	1000
macro avg	0.99	0.98	0.98	1000
weighted avg	0.99	0.98	0.98	1000

Figure 4: Classification report on Prediction Model

Excelling in recognizing “Happy” and “Neutral” emotions, though some misclassification occurred between “Fear” and “Disgust” due to visual similarities. The CNN model, trained on ECG signals, reached 98.5% accuracy, effectively distinguishing “Happy” and “Sad” but occasionally confusing “Neutral” and “Sad” due to physiological similarities. After combining features from both models, the integrated system achieved a higher accuracy of 93.3%, demonstrating substantial improvements, particularly in distinguishing difficult emotions.

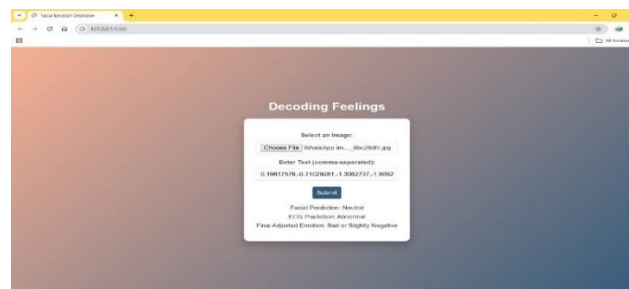


Figure 5: Final Output

The AUC score of 0.97 highlighted its strong discriminative capability, and the updated confusion matrix showed reduced misclassification. Performance metrics, including precision, recall, and F1-score, confirmed the benefits of multi-modal fusion. The system’s robustness and reliability make it suitable for applications such as health monitoring and adaptive AI systems.

VI. DISCUSSION

- Multimodal Emotion Recognition: The integration of facial expressions (visual) and ECG signals (physiological) captures both external and internal emotional states, providing a more comprehensive understanding of emotions compared to unimodal approaches.

- **Innovative Deep Learning Framework:** The use of MobileNetV2 for efficient facial feature extraction and a specialized CNN for analyzing ECG signals ensures real-time emotion recognition with high accuracy.
- **Attention-Based Data Fusion:** The attention mechanism optimally combines the strengths of visual and physiological data, enhancing the model's ability to classify emotions more accurately.
- **Comprehensive Emotion Classification:** The model accurately classifies five distinct emotions—sadness, happiness, fear, neutrality, and disgust—demonstrating robust emotion recognition [9] capabilities.
- **Real-World Applicability:** The system is scalable and designed for real-time performance, making it ideal for applications such as mental health tools, adaptive interfaces, and emotion-aware devices.
- **Significant Performance Improvement:** The proposed method outperforms traditional unimodal approaches, offering superior accuracy in emotion classification.
- **Potential Applications:** The framework is suitable for various use cases, including personalized learning systems, mental health support[26], and adaptive user interfaces for human-computer interaction.

REFERENCES

- [1] Li, S., Deng, W.: Deep facial expression recognition: A survey. *IEEE transactions on affective computing* 13(3), 1195–1215 (2020)
- [2] Zaheer, R., Ahmad, I., Habibi, D., Islam, K.Y., Phung, Q.V.: A survey on artificial intelligence-based acoustic source identification. *IEEE Access* 11, 60078–60108 (2023)
- [3] Zhang, Y., Cheng, C., Zhang, Y.: Multimodal emotion recognition using a hierarchical fusion convolutional neural network. *IEEE access* 9, 7943–7951 (2021)
- [4] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., Zong, Y.: A survey of deep learning- based multimodal emotion recognition: Speech, text, and face. *Entropy* 25(10), 1440 (2023)
- [5] Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech: a review. *International journal of speech technology* 15, 99–117 (2012)
- [6] Xu, e.a. T.: Emotion recognition from physiological and speech signals for mental health monitoring. *IEEE Transactions on Affective Computing* (2024)
- [7] Yao, e.a. Y.: Emotion detection using physiological signals for healthcare applications. *Sensors and Actuators A: Physical* (2022) <https://doi.org/10.1016/j.sna.2022.113785>
- [8] Zhao, e.a. Y.: Emotion detection from multimodal data for personal health systems. *Journal of Biomedical and Health Informatics* (2024)
- [9] Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and Vision Computing* 133, 104676 (2023)
- [10] Zhang, e.a. Z.: Facial expression recognition with deep learning: A review. *IEEE Transactions on Affective Computing* (2020) <https://doi.org/10.1109/TAFFC.2019.2931052>
- [11] Ali, e.a. A.: Emotion detection using convolutional neural networks for real-time applications. *Journal of Machine Learning Research* (2024)
- [12] Ranan, e.a. P.: Deep learning techniques for emotion recognition in multimodal systems. *Proceedings of the IEEE International Conference on Neural Networks* (2024)