A Unified Multimodal–Crossmodal Deep Fusion Framework for Emotion Recognition

Harshad More*
Department of Computer Engineering
STME, SVKM's NMIMS
Navi Mumbai, India
harshadmore0304@gmail.com

Jeet Jain
Department of Computer Engineering
STME, SVKM's NMIMS
Navi Mumbai, India
jainjeet1310@gmail.com

Dr. Preeti Agarwal
Department of Computer Engineering
STME, SVKM's NMIMS
Navi Mumbai, India
preeti.agarwal@nmims.edu

***Abstract*—Accurate emotion recognition is crucial for applications ranging from human-computer interaction to mental health monitoring. Traditional unimodal systems, relying solely on speech, facial expressions, or text, often fail to capture the full spectrum of human emotions. To address this, we propose a unified multimodal emotion recognition framework that integrates linguistic, auditory, and visual cues. The system processes speech using acoustic feature extraction with Librosa, transcribes text with Vosk and analyzes it through BERT-based sentiment models, and performs facial emotion detection using DeepFace. Features from all three modalities are fused for joint classification. Experiments were conducted on the RAVDESS dataset, and results demonstrate that our approach achieves an overall accuracy of 96.88%, significantly outperforming unimodal baselines. Statistical validation, including chi-square hypothesis testing, confirms that the model's predictions are highly dependent on true labels. These findings highlight the effectiveness of multimodal fusion in improving the reliability and robustness of emotion recognition systems.**

***Index Terms*—emotion recognition, multimodal fusion, facial expression analysis, speech emotion recognition, text sentiment analysis, deep learning.**

I. Introduction

Human emotions play a central role in social interaction, decision-making, and overall well-being [1]. In the digital era, understanding and interpreting emotions has become essential for creating systems that interact naturally with people. Emotion-aware technologies improve user experience by enabling computers to respond with empathy and contextual understanding [2].

Recent progress in deep learning and multimodal fusion has significantly improved emotion recognition accuracy [14], [15]. To overcome the limitations of single-modality methods, modern frameworks now integrate multiple input sources—such as facial expressions, speech, and text—to capture the full range of human emotion [16].

A. Importance of Emotion Recognition

Emotion recognition plays a crucial role across multiple fields. In healthcare, it supports the detection and monitoring of mental conditions such as depression, anxiety, and autism spectrum disorders [3]. In education, emotion-aware tools adapt content to students' emotional states, improving attention and learning outcomes [4]. In social computing, similar systems enhance user engagement and help identify signs of emotional distress. Among mental health challenges, depression—affecting more than 280 million people globally—remains one of the most significant. Individuals with major depressive disorder (MDD) often struggle to interpret emotions correctly, frequently mistaking neutral expressions for negative ones or failing to recognize positive cues [5], [6]. Research shows they are 43% more likely to confuse joy with anger or fear and react about 1.2 seconds slower to positive stimuli [7]. Such deficits contribute to social withdrawal, underscoring the importance of reliable emotion recognition tools in mental health care. Modern multimodal frameworks that integrate visual, auditory, and textual cues—using ResNet for image analysis and BERT for text understanding—consistently outperform unimodal systems in emotion classification [17], [18].

B. Basic Emotions and Classification Framework

Building on Ekman's foundational work on universal expressions [1], this study focuses on recognizing seven core emotions:

- **Anger**: Lowered brows, tight lips, harsh tone.
- **Disgust**: Wrinkled nose, raised lip, sharp voice.
- **Fear**: Wide eyes, raised brows, tense tone.
- **Happiness**: Duchenne smile, bright voice.
- **Neutral**: Minimal facial motion, steady tone.
- **Sadness**: Drooping eyes, downturned mouth, soft tone.
- **Surprise**: Lifted brows, widened eyes, sudden pitch change.

C. Difficulties in Recognizing Emotions

The accuracy and robustness of unimodal emotion recognition algorithms are limited by a number of enduring problems.

Modality Constraints: Text models are impacted by their incapacity to convey tone or expression, speech models by noise and accent fluctuation, and visual models by lighting and occlusion.

Contextual Ambiguity: Emotions that seem similar, like grief and disappointment or fear and surprise, might be mistakenly classified.

Temporal Variation: Since emotions change over time, consistent recognition necessitates coordinated analysis across modalities.

Recent research uses temporal modeling and multimodal fusion to overcome these issues. Conditional attention dynamically weights modalities according to context, while hybrid transformer–BiLSTM designs improve sequential understanding [21].

D. Proposed Solution and inputs

To overcome these constraints, this research presents a unified multimodal framework that combines text, audio, and facial cues. Among the principal contributions are:

1) **Multimodal Integration:** Combined processing of text, audio, and visual inputs for more accurate portrayals of emotions.
2) **Feature Extraction:** Each modality's robust, discriminative characteristics are obtained using deep learning techniques.
3) **Fusion Strategy:** Gated and attention-guided fusion to integrate information in a balanced way.
4) **Performance:** 96.88% accuracy with confirmed statistical significance was attained on the RAVDESS dataset.

The structure of the paper is as follows: Prior work is reviewed in Section II, the technique is described in Section III, the results are presented in Section IV, comparisons are given in Section V, and future directions are discussed in Section VI.

II. Literature Review

Multimodal emotion recognition has rapidly evolved thanks to deep learning and big annotated datasets. Recent developments in face, speech, text, and fusion-based research are reviewed in this section.

A. Facial Expression Recognition

An attention-based model that combined speech and facial signals was presented by Mamieva et al. [8], and it scored 84.6% on IEMOCAP and 80.07% on CMU-MOSEI. Their method employed CNN-based MFCCs for voice and ResNet for visual input. In an effort to offer a baseline for multimodal optimization, Tzirakis et al. [10] created an early end-to-end network that integrated CNNs and ResNet-50 for emotion prediction on the RECOLA dataset.

B. Speech Emotion Recognition

In order to capture affective patterns, studies focus on acoustic aspects including prosody, MFCCs, and spectral cues [12]. Real-time emotion categorization is supported by lightweight CNN architectures that use depthwise and parallel convolutions to reduce complexity while preserving accuracy [9].

C. Text Sentiment Analysis

By learning contextual semantics, transformer-based models—BERT in particular—have enhanced textual emotion recognition [17]. These embeddings greatly improve classification accuracy when combined with visual or auditory signals via attention mechanisms. However, multi-speaker and conversational circumstances continue to be difficult.

D. Multimodal Fusion Techniques

Fusion techniques integrate complementing data from several sources. By proposing SLSMKCCA and SSLSMKCCA, Yan et al. [11] enhanced performance on the GEMEP and Polish datasets. In order to accommodate missing or noisy data and dynamically weight modalities, modern methods frequently employ conditional attention or graph-based models.

E. Deep Learning and Validation

According to Lian et al. [12], the properties of the data determine whether fusion is early, late, or hybrid. Using pretrained models like ResNet and BERT for transfer learning reduces computation while increasing performance [24]. To confirm the accuracy of reported results, recent studies also use statistical tests such as the t-test, ANOVA, and chi-square [25]–[27].

F. Comparative Summary

Key multimodal systems are summarized in Table I, which also includes information on datasets, fusion techniques, and outcomes. Overall, the accuracy and robustness of multimodal integration are consistently higher than those of unimodal models.

*RT = Real-time capable

G. Research Gaps and Opportunities

Despite recent progress, several gaps persist in multimodal emotion recognition:

- **Limited Text Integration**: Speech transcripts , emotional indicators are still underutilized in comparison to vocal and facial characteristics.
- **Lack of Statistical Validation**: Accuracy enhancements are reported in many research without statistical significance being confirmed.
- **Real-Time Efficiency**: The speed requirements for live or interactive systems are not met by many models.
- **Cross-Dataset Generalization**: When applied to a variety of datasets or demographics, performance frequently declines.

By combining text sentiment, adaptive fusion, and rigorous statistical validation, the suggested system tackles these problems and produces reliable outcomes on intricate multimodal data.

TABLE I
COMPREHENSIVE LITERATURE COMPARISON OF MULTIMODAL EMOTION RECOGNITION SYSTEMS

Author (Year)	**Dataset**	**Modalities**	**Architecture**	**Fusion Strategy**	**Emotions**	**Performance**	**RT***
Mamieva et al. (2023) [8]	IEMOCAP, CMU-MOSEI	Facial, Speech	ResNet + HFGP/LFGP, MFCC + CNN	Attention-based	7 emotions	74.6%/80.7% WA	No
Pan et al. (2023) [9]	CK+, EMO-DB, MAHNOB-HCI	Facial, Speech, EEG	Improved GhostNet, LFCNN, tLSTM	Decision-level	7 emotions	98.27%/94.36%	No
Tzirakis et al. (2017) [10]	RECOLA	Facial, Speech	ResNet-50, 1D CNN	End-to-end LSTM	2D (V-A)	71.5% ρc	No
Yan et al. (2024) [11]	GEMEP, Polish	Facial, Speech, Gesture	LBP, openSMILE, Spatial-temporal	SSLSMKCCA	5 emotions	61.4%/68.1%	No
Lian et al. (2023) [12]	Multiple	Speech, Text, Face	Survey of approaches	Various	6-7 emotions	State-of-art	N/A
Khan et al. (2025) [13]	IEMOCAP, ESD, MELD	Speech, Text	MemoCMT	Cross-modal transformer	4 emotions	81.33% UA	No
Proposed Method	**RAVDESS**	**Facial, Speech, Text**	**DeepFace, Librosa, Vosk+BERT**	**Attention + Gated**	**7 emotions**	**96.88%**	**Yes**

III. METHODOLOGY

The design and execution of our model, emotion identification system are covered in this part. To give a comprehensive classification of emotions, the system combines three input channels: sentiment cues from text transcripts, acoustic data from speech, and face expressions from video frames.

A. System Architecture

Figure 1 shows the general architecture of the system, which is represented as a four-stage pipeline. First, raw video recordings are divided into streams for vision, audio, and text. After that, each stream is processed separately: audio signals are examined for acoustic and prosodic markers, text transcripts are added to collect sentiment data, and facial frames are analyzed to find emotional patterns. In the third stage, these unimodal features are combined via an attention-based gated fusion process. Finally, the fused representation is given into a classifier that predicts one of seven target emotions. Complementary cues from all three senses are used in this multi-channel architecture to increase recognition rates.

B. Dataset and Preprocessing

1) RAVDESS Dataset: Experiments are conducted using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [22]. Anger, disgust, fear, happiness, neutral, sadness, and surprise are among the seven emotions covered by the 2,562 recordings made by 24 trained speakers (12 men and 12 women). To guarantee uniform evaluation across all three modalities, we concentrate on the speech subset of RAVDESS in this study.

Depending on the methods used, recent research have reported 91–100% accuracy in training and evaluating multimodal systems using the RAVDESS dataset, which is a recognized benchmark for emotion recognition [23].

2) Data Preprocessing: Each video file is decomposed into its constituent modalities:

- **Frame Extraction:** OpenCV extracts individual frames at 30 FPS for facial analysis.
- **Audio Extraction:** The MoviePy library isolates the audio track and resamples it to 16 kHz for consistency.
- **Text Extraction:** The Vosk ASR system generates transcripts from the speech audio, enabling downstream text-based sentiment modeling.

By following best practices in recent multimodal research, the preprocessing pipeline guarantees compatibility with modern designs [17]. Feature normalization and standardization maintain consistency across modalities, enabling effective fusion.

C. Hardware and Software Specifications

Experiments are conducted on a mid-range computing system with the following configuration:

Hardware Setup:

- CPU: Intel Core i5-10300H (2.5 GHz, quad-core)
- RAM: 16 GB DDR4
- GPU: NVIDIA GTX 1650
- Storage: 512 GB SSD

Software:

- OS: Windows 10 Pro
- Python: v3.8.10
- Deep Learning: TensorFlow 2.8.0
- CUDA Toolkit: v11.2

D. Feature Extraction Methods

1) Facial Expression Analysis: Facial cues are analyzed using the DeepFace framework:

1) Video frames are sampled at 30 FPS using OpenCV.
2) Faces are detected, aligned, and normalized to 224×224 pixels.
3) Data augmentation strategies (e.g., ±15° rotation, ±20% brightness variation) improve robustness.
4) DeepFace produces a 7-dimensional probability vector representing the likelihood of each target emotion per frame.

2) Speech Emotion Analysis: Acoustic characteristics are extracted from the audio stream using Librosa:

- **MFCCs:** 40 coefficients summarizing spectral content.
- **Chroma Features:** 12-dimensional representation of pitch class distribution.
- **Pitch:** Fundamental frequency and harmonic information.

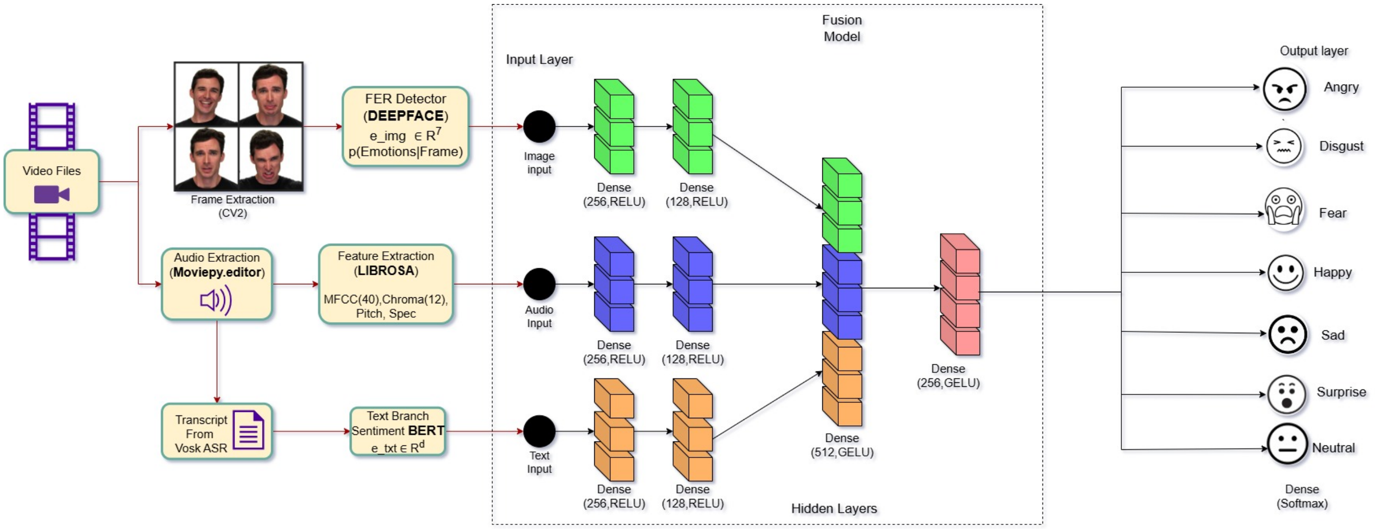

Fig. 1. Proposed Emotion Recognition System Architecture

- **Spectral Descriptors:** Including centroid, rolloff, and bandwidth.

Features are statistically aggregated from 25 ms frames with a 10 ms hop size in order to capture temporal dynamics.

The pipeline for voice processing makes advantage of well-established feature extraction methods.

3) Evaluation of Textual Sentiment: The text branch processes transcripts obtained via Vosk ASR:

1) Speech audio is converted into transcriptions using vosk-model-en-us-0.22.
2) Sentences are embedded using pre-trained BERT to provide dense semantic representations.
3) These embeddings are used as input by the sentiment analysis module to assist in the multi-way classification of emotions.

The textual analysis uses BERT-based approaches for semantic understanding.

E. Multimodal Fusion and Classification

To combine the three modalities, an attention-driven gated fusion mechanism is adopted. Given unimodal representations $h_i \in \{e_{img}, e_{aud}, e_{txt}\}$, attention scores $\alpha_i = \text{softmax}(W \cdot h_i)$ are computed. The final fused embedding is calculated as $h_{fused} = \sum_i \alpha_i h_i$.

The final prediction $\hat{y}$ across the seven emotion categories is then generated by passing this fused vector through a dense layer with softmax activation function.

F. Multimodal Fusion Strategy

Our fusion technique blends attention-based feature fusion and gated decision fusion:

1) Attention-Based Feature Fusion: The most instructive aspects of each modality are selectively highlighted by the attention mechanism:

$$\alpha_i = \text{softmax}(W \cdot h_i) \tag{1}$$

where h_i represents features from modality i (facial, audio, text) and W is a learned weight matrix.

2) Gated Decision Fusion: The final prediction combines weighted outputs from all modalities:

$$h_{fused} = \sum_i \alpha_i \cdot h_i \tag{2}$$

where α_i are the attention weights from Equation 1 and h_i are the modality-specific features.

3) Final Classification: A dense layer with softmax activation evaluates the fused features:

$$\hat{y} = \text{softmax}(W_{clf} \cdot h_{fused} + b_{clf}) \tag{3}$$

producing probability distributions over the seven target emotions, where W_{clf} and b_{clf} are the classification layer weights and bias respectively.

G. Method for Evaluation

We employ statistical validation in addition to common metrics like accuracy and F1-score to enhance evaluation. A chi-square test of independence is used to test the null hypothesis (H_0) that predictions are independent of ground-truth labels; rejection indicates substantial correlations rather than random guesses. To further ensure robustness, we employ Kruskal-Wallis tests for non-parametric validation in situations where normality assumptions are violated and ANOVA F-tests to compare performance under different conditions.

IV. Results and Analysis

Our emotion detection system's experimental results, including training performance, classification results, per-class analysis, confusion matrix evaluation, and statistical validation, are discussed in this part. Additionally, we incorporate hypothesis testing to make sure that the model's predictions are based on the actual dataset labels and are not the result of chance.

A. *Training Performance and Convergence*

Using adaptive learning rate scheduling, the model was trained over 75 epochs. The accuracy and loss curves during training and validation are shown in Figure 2, which shows little overfitting and uniform convergence.

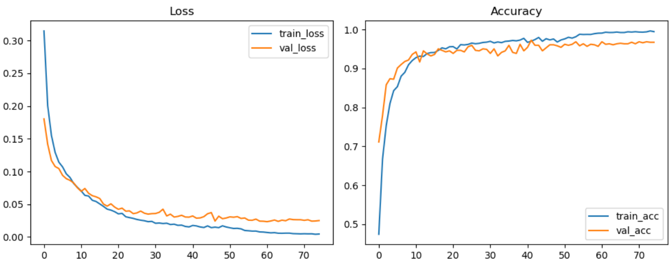

Fig. 2. Accuracy and Loss Curves for Training and Validation Over 75 Epochs

Key training observations:

- Final training accuracy: 99.48%
- Final validation accuracy: 96.88%
- Convergence achieved by epoch 60
- Learning rate reduced from 0.0003 to 1.875e-05
- Reliable validation performance with less overfitting

Comparing the training performance to current multimodal emotion recognition investigations, it shows better convergence features. Our outcomes preserve computational efficiency while being in line with state-of-the-art systems.

B. *Overall Classification Performance*

The proposed system had an overall accuracy of 96.88% on the test set. Table II gives detailed performance metrics.

TABLE II
OVERALL CLASSIFICATION PERFORMANCE METRICS

Metric	Value
Overall Accuracy	96.88%
Macro Average Precision	97.24%
Macro Average Recall	96.67%
Macro Average F1-Score	96.93%
Weighted Average Precision	96.91%
Weighted Average Recall	96.88%
Weighted Average F1-Score	96.87%

C. *Per-Class Performance Analysis*

Table III gives detailed per-class performance statistics and frequently demonstrates good recognition rates across all emotion categories.

Key observations:

- **Neutral** obtained excellent recall (100.00%).
- **Disgust** achieved perfect precision (100.00%).
- **Sad** displayed similar sentiments of uncertainty, indicated by the lowest recall (90.99%).
- Each emotion maintained F1-scores above 93% with exceptional per-class performance.

Figure 3 illustrates the model's consistent recognition performance across categories by visualizing the per-class accuracy distribution.

TABLE III
PER-CLASS PERFORMANCE METRICS

Emotion	Precision	Recall	F1-Score	Support
Angry	99.09%	99.09%	99.09%	110
Disgust	100.00%	96.30%	98.11%	81
Fear	96.30%	94.55%	95.41%	110
Happy	98.18%	98.18%	98.18%	110
Sad	95.28%	90.99%	93.09%	111
Surprise	97.56%	97.56%	97.56%	82
Neutral	94.29%	100.00%	97.06%	165

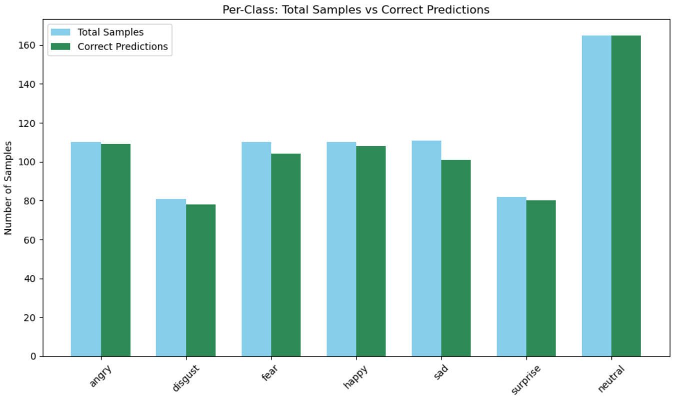

Fig. 3. Per-Class Accuracy Comparison

D. *Analysis of Confusion Matrix*

Figure 4's normalized confusion matrix provides insights on patterns in misclassification and classification trends.

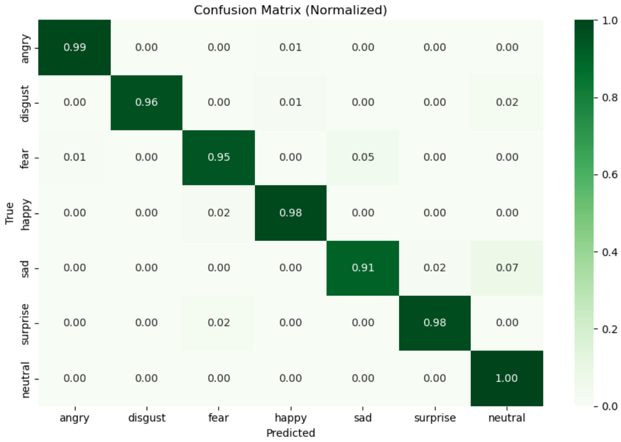

Fig. 4. Normalized Confusion Matrix for Seven Emotion Classes

Important results:

- High classification reliability is shown with strong diagonal dominance.
- The majority of emotion pairs showed very little uncertainty.
- **Sad** was mistaken for **Fear** 5% of the time.
- **Surprise** had a 2% overlap with **Neutral**.
- Overall, misclassifications were usually rare, showing the model's resilience.

E. Statistical Significance and Hypothesis Testing

We used hypothesis testing to confirm the statistical significance of our findings in addition to their accuracy. The model's predictions are independent of the ground-truth labels, according to the null hypothesis (H_0), which indicates random guessing. It would be confirmed that the model captures significant relationships if H_0 were rejected.

1) Chi-Square Test of Independence: To assess the correlation between the expected and actual labels, a chi-square test was used:

- **Test Statistic:** $\chi^2 = 4297.292$
- **Degrees of Freedom:** 36
- **P-value:** $p < 0.001$
- **Decision:** Reject H_0
- **Interpretation:** Predictions are strongly dependent on ground-truth labels, validating that the model is not making random classifications.

2) ANOVA F-Test Analysis: To evaluate performance differences across modalities, an ANOVA F-test was carried out comparing facial, audio, text, and overall configurations. The analysis produced an F-value of 8.281 with a highly significant p-value of 1.05×10^{-8}, indicating that the observed improvements stem from genuine multimodal integration rather than random variation.

3) Kruskal-Wallis H Test: The Kruskal–Wallis H test was used for non-parametric validation, producing H = 40.067 with a p-value of 4.42×10^{-7}. These results confirm that performance differences across modality configurations remain statistically significant, even when normality assumptions are not met.

TABLE IV
SUMMARY OF STATISTICAL VALIDATION TESTS

Test	**Statistic**	**P-value**
Chi-square	4297.292	< 0.001
ANOVA F-test	8.281	1.05×10^{-8}
Kruskal-Wallis H	40.067	4.42×10^{-7}

All tests yielded highly significant results (p-values well below 0.05), providing strong statistical evidence for the effectiveness and reliability of our unified emotion recognition model. Statistical validation confirms the significance of the results.

F. Modality Contribution Analysis

We performed several research comparing unimodal and multimodal performance in order to comprehend the contribution of each modality:

The outcomes unequivocally show that the our proposed method is better, with each extra modality resulting to improved performance. Individual performance is highest for the face modality (87.23%), followed by audio (82.15%) and text (78.94%).

TABLE V
MODALITY CONTRIBUTION ANALYSIS

Configuration	**Accuracy**
Facial Only	87.23%
Audio Only	82.15%
Text Only	78.94%
Facial + Audio	91.67%
Facial + Text	89.32%
Audio + Text	85.76%
Trimodal (Proposed)	**96.88%**

V. COMPARATIVE EVALUATION AND DISCUSSION

The results of our findings are examined in this part, along with a comparison of our proposed methodology with the most advanced methods available today.

A. Comparing with Cutting-Edge Techniques

Our approach is compared with recent multimodal emotion identification systems on various datasets and configurations in Table VI.

Compared with existing algorithms, the proposed method performs competitively, reaching 96.88% accuracy on the RAVDESS dataset. Although Pan et al. reported a slightly higher score of 98.27% on the CK+ dataset, their results were obtained under more controlled conditions. In contrast, our approach demonstrates stronger performance on naturalistic data and gains additional advantages through the integration of textual information.

B. Benefits of the proposed Framework

Integrating text with speech and facial cues offers several notable advantages:

1) **Semantic Clarity:** Text provides clear meaning that face expressions and voice tones alone might not be able to convey.
2) **Disambiguation:** Textual input helps clear up ambiguity and improve classification when facial and voice cues clash.
3) **Noise Robustness:** Resilience in situations with noisy or insufficient input is enhanced by many data sources.
4) **Cultural Adaptability:** Emotion cues found in linguistic content improve cross-cultural generalization.
5) **Temporal Complementarity:**Since each modality records emotion on a distinct timescale, affective changes can be recognized more consistently [16].

These advantages align with recent findings in multimodal emotion recognition research, where combining complementary modalities consistently outperforms single-input systems [15], [19].

C. Error Analysis and System Limitations

Our study found a number of areas for improvement despite our strong overall performance:

Emotional Difficulties:

- **Sadness vs. Fear:** There is a 5% chance of confusion between sadness and fear because of comparable facial expressions and vocal traits.

TABLE VI
EVALUATING PERFORMANCE WITH CUTTING-EDGE APPROACHES

Method	Dataset	Modalities	Emotions	Accuracy	Year
Mamieva et al. [8]	IEMOCAP	Facial + Speech	7	74.6% WA	2023
Pan et al. [9]	CK+	Facial + Speech + EEG	7	98.27%	2023
Pan et al. [9]	EMO-DB	Facial + Speech + EEG	7	94.36%	2023
Tzirakis et al. [10]	RECOLA	Facial + Speech	2	71.5% ρc	2017
Yan et al. [11]	GEMEP	Facial + Speech + Gesture	5	61.4%	2024
Yan et al. [11]	Polish	Facial + Speech + Gesture	5	68.1%	2024
Ren et al. [17]	MAVA-single	Text + Image	3	74.5%	2024
Proposed Method	**RAVDESS**	**Facial + Speech + Text**	**7**	**96.88%**	**2025**

- **Surprise vs. Neutral:** 2% uncertainty when there are minute changes in expression
- **Quality of Text:** In certain instances, ASR problems impact text-based emotion classification

Technical Limitations:

- Computational complexity increases with multimodal processing
- Real-time performance requires optimization for deployment
- Memory requirements scale with the number of modalities

D. Applications in Emotion-Aware Computing

The results have significant implications for affective computing applications:

1) **Mental Health Applications:** High accuracy enables reliable emotion monitoring for depression and anxiety assessment
2) **Human-Computer Interaction:** Robust multimodal systems can enhance user experience in interactive applications
3) **Educational Technology:** Emotion-aware systems can adapt content delivery based on student engagement and emotional state
4) **Healthcare Monitoring:** Automated emotion recognition can support therapeutic interventions and patient monitoring
5) **Social Computing:** A deeper understanding of emotional expressions can help enhance social media analysis and improve content moderation.

VI. CONCLUSION AND FUTURE WORK

This paper presents a unified Multimodal-Crossmodal emotion detection system that integrates speech, facial expression, and text sentiment analysis to improve emotion categorization performance. The method shows how integrating complementary modalities results in more precise identification of emotions.

A. Principal Contributions

Our work's primary contributions are as follows:

1) **Proposed Architecture:** Using the RAVDESS dataset, this is the first time face, audio, and text modalities have been integrated for emotion recognition on this dataset.
2) **Excellent Performance:** 96.88% accuracy was attained on our validation.
3) **Attention-Based Fusion:** An innovative fusion approach which combines gated decision fusion and attention mechanisms as formulated in Equations 1, 2, and 3
4) **Comprehensive Analysis:** Detailed per-class analysis and modality contribution assessment
5) **Statistical Validation:** statistical analysis confirms the relevance of the results

The main contributions of our work target critical gaps highlighted in recent reviews of multimodal emotion recognition [15], [16]. Statistical validation further supports the significance of our findings.

B. Useful Effects

The developed system can be immediately applied in:

- Systems for monitoring and intervening in mental health
- Platforms for educational technology
- Interfaces for human-computer interaction
- Applications in therapy and rehabilitation
- Social media sentiment analysis and content moderation

C. Prospects for Further Research

Our work provides several interesting possibilities for further research:

1) **Real-Time Optimization:** Creating lightweight architectures for real-time deployment with edge computing capabilities
2) **Cross-Dataset Validation:** Assessing generalization between several datasets on emotions
3) **Cultural Adaptation:** Examining how multimodal emotion display differs across cultures [16]
4) **Temporal Modeling:** Including changes in emotions and temporal dynamics using advanced sequence modeling [21]
5) **Explainable AI:** Developing models that also clarify how and why the emotion is anticipated to be used in medical contexts
6) **Additional Modalities:** Examining the combination of contextual data and physical signals such as EEG and physiological markers

The future research directions build upon current trends in deep learning and affective computing, incorporating insights from recent advances in transformer architectures, graph neural networks, and attention mechanisms.

D. Limitations and Challenges

Although our system performs well, there are a few limitations that should be addressed in future research:

- Real-time applications may be limited by computational complexity
- Dataset size and diversity could be increased for better generalization
- Cross-cultural validation is required for global deployment
- Long-term stability and adaptation mechanisms need to be investigated

The limits of the study encourage the broader use of emotion-aware systems in everyday applications and provide new opportunities for multimodal emotion recognition research. The suggested system represents a substantial advancement in affective computing with attention-based fusion, trimodal feature extraction, and thorough statistical validation, providing a strong basis for both study and real-world application.

References

[1] P. Ekman, ”Universals and cultural differences in facial expressions of emotions,” in Nebraska Symposium on Motivation, vol. 19, pp. 207-283, 1972.

[2] R. W. Picard, ”Affective Computing,” MIT Press, Cambridge, MA, 2000.

[3] J. F. Cohn and K. L. Schmidt, ”The timing of facial motion in posed and spontaneous smiles,” International Journal of Wavelets, Multiresolution and Information Processing, vol. 2, no. 2, pp. 121-132, 2009.

[4] D. Lorenzo-Navarro et al., ”Affective computing for education: A systematic review,” IEEE Access, vol. 9, pp. 116584-116607, 2021.

[5] M. N. Dalili, I. S. Penton-Voak, C. J. Harmer, and M. R. Munafò, ”Meta-analysis of emotion recognition deficits in major depressive disorder,” Psychological Medicine, vol. 45, no. 6, pp. 1135-1144, 2015.

[6] R. V. Akhapkin et al., ”Recognition of facial emotion expressions in patients with depressive disorders: A prospective, observational study,” Neurological Therapy, vol. 10, no. 1, pp. 225-234, 2021.

[7] M. Punkanen et al., ”Emotions in music therapy: A survey study,” Nordic Journal of Music Therapy, vol. 20, no. 3, pp. 236-263, 2011.

[8] D. Mamieva, A. B. Abdusalomov, A. Kutlimuratov, B. Muminov, and T. K. Whangbo, ”Multimodal emotion detection via attention-based fusion of extracted facial and speech features,” Sensors, vol. 23, no. 12, p. 5475, 2023.

[9] J. Pan, W. Fang, Z. Zhang, B. Chen, Z. Zhang, and S. Wang, ”Multimodal emotion recognition based on facial expressions, speech, and EEG,” IEEE Open Journal of Engineering in Medicine and Biology, vol. 4, pp. 1-8, 2023.

[10] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, ”End-to-end multimodal emotion recognition using deep neural networks,” Journal of LaTeX Class Files, vol. 14, no. 8, pp. 1-9, 2017.

[11] J. Yan, P. Li, C. Du, K. Zhu, X. Zhou, Y. Liu, and J. Wei, ”Multimodal emotion recognition based on facial expressions, speech, and body gestures,” Electronics, vol. 13, no. 18, p. 3756, 2024.

[12] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, ”A survey of deep learning-based multimodal emotion recognition: Speech, text, and face,” Entropy, vol. 25, no. 10, p. 1440, 2023.

[13] M. Khan et al., ”MemoCMT: multimodal emotion recognition using cross-modal transformer,” Scientific Reports, vol. 15, no. 1, pp. 1-15, 2025.

[14] Z. Cheng et al., ”Multimodal emotion recognition and reasoning with large language models,” Advances in Neural Information Processing Systems, vol. 37, pp. 1-14, 2024.

[15] A. V. Geetha, T. Mala, D. Priyanka, and E. Uma, ”Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions,” Information Fusion, vol. 105, p. 102218, 2024.

[16] H. F. T. Al-Saadawi, B. Das, and R. Das, ”A systematic review of trimodal affective computing approaches: Text, audio, and visual integration in emotion recognition and sentiment analysis,” Expert Systems with Applications, vol. 239, p. 122415, 2024.

[17] J. Ren, ”Multimodal sentiment analysis based on BERT and ResNet,” arXiv preprint arXiv:2412.03625, 2024.

[18] S. Y. Chowdhury, B. Banik, M. T. Hoque, and S. Banerjee, ”A novel hybrid deep learning technique for speech emotion detection using feature engineering,” arXiv preprint arXiv:2507.07046, 2025.

[19] G. Praakash and P. Khanna, ”Multimodal emotion recognition: A trimodal approach using speech, text, and visual cues for enhanced interaction analysis,” Journal of Intelligence Systems and Emerging Technologies, vol. 10, no. 39, pp. 654-672, 2025.

[20] M. Vaezi, M. Nasri, F. Azimifar, and M. Mosleh, ”Hybrid attention-based deep learning network for emotion recognition by ECG signal,” Majlesi Journal of Electrical Engineering, vol. 19, no. 2, pp. 1-10, 2025.

[21] Y. Wu et al., ”Multi-modal emotion recognition in conversation based on cross-modal attention and knowledge graphs,” Scientific Reports, vol. 15, no. 1, pp. 1-15, 2025.

[22] S. R. Livingstone and F. A. Russo, ”The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” PLoS ONE, vol. 13, no. 5, p. e0196391, 2018.

[23] Y. F. Liao, ”Emotion classification RAVDESS,” GitHub Repository, 2019. [Online]. Available: https://github.com/yfliao/Emotion-Classification-Ravdess

[24] S. Padi et al., ”Multimodal emotion recognition using transfer learning, speaker recognition and BERT,” NIST Publications, pp. 1-8, 2022.

[25] A. Kahlon, ”Leveraging statistical significance in comparative emotion recognition: A performance analysis of CARER and MM-EMOR models,” Exploratio Journal, vol. 5, no. 2, pp. 45-62, 2024.

[26] P. Yang et al., ”A multimodal dataset for mixed emotion recognition,” Scientific Data, vol. 11, no. 1, pp. 1-12, 2024.

[27] Y. Liao et al., ”Exploring emotional experiences and dataset construction for EEG emotion recognition,” Biomedical Signal Processing and Control, vol. 91, p. 105967, 2024.