CHAPTER 8

Multimodal AI for Emotion Recognition: Integrating Speech, Text, and Facial Expressions

Mr. Vorem Kishore

Assistant Professor, Department of Computer Science and Engineering-AIML and IoT, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, Telangana, India.

Email: kishore_v@vnrvjiet.in

https://doi.org/10.58599/GSE.2025.081208

Abstract: Emotion recognition has become a pivotal area of research in human-computer interaction, artificial intelligence, and affective computing. While unimodal approaches have shown promise, they are often limited by the inherent ambiguity and subtlety of human emotional expression. This chapter explores the paradigm of Multimodal Artificial Intelligence (AI) for emotion recognition, a more robust approach that integrates information from multiple sources—specifically speech, text, and facial expressions. We delve into the foundational concepts of multimodal systems, from data preprocessing and feature extraction to advanced fusion techniques. A comprehensive literature review is presented, highlighting seminal works and state-of-the-art models that have shaped the field. We then propose a novel hybrid deep learning framework that leverages Convolutional Neural Networks (CNNs) for spatial feature extraction from facial and speech data, and Long Short-Term Memory (LSTM) networks for capturing temporal dependencies. The chapter details the proposed methodology, including the architecture, feature extraction pipelines for each modality, and a hybrid fusion strategy designed to maximize inter-modal correlations. An extensive Results and Discussions section presents simulated experimental results on benchmark datasets, demonstrating the superiority of the multimodal approach over unimodal systems. We analyze performance metrics, including accuracy, F1-score, and confusion matrices, and compare different fusion strategies. The chapter concludes with a summary of key findings, a discussion of the challenges and limitations of current methods, and an outlook on future research directions in multimodal emotion recognition, paving the way for more empathetic and intelligent applications.

Keywords: Multimodal Emotion Recognition; Feature Fusion; Convolutional Neural Networks; Long Short-Term Memory; Human–Computer Interaction.

1. Introduction

Human communication is a rich tapestry woven from verbal and non-verbal cues. The words we speak, the tone of our voice, and the expressions on our faces all contribute to the emotional message we convey. For artificial intelligence to achieve true human-like understanding and interaction, it must be capable of perceiving and interpreting this complex, multimodal emotional landscape. Emotion Recognition is the task of automatically identifying human emotions, a capability that promises to revolutionize fields ranging from mental healthcare and customer service to education and entertainment [1]. Early research in this domain predominantly focused on unimodal systems, analyzing one modality at a time. For instance, facial expression analysis has used computer vision to classify emotions from static images or video frames [2]. Similarly, speech emotion recognition has analyzed acoustic features like pitch, intensity, and spectral content to infer emotional states [3]. Text-based sentiment analysis, on the other hand, has relied on natural language processing (NLP) to determine the emotional polarity of written content. However, these unimodal systems face significant limitations. A single modality can be ambiguous; a smile can be genuine or sarcastic, and the phrase “that’s great” can be sincere or ironic. The true emotional context often lies in the interplay between these different channels. This limitation has given rise to Multimodal Emotion Recognition, an approach that integrates data from multiple sources to form a more holistic and accurate understanding of human emotion. By combining information from speech, text, and facial expressions, AI systems can disambiguate conflicting signals and capture the nuances of emotional expression that are lost in a single modality. For example, a system might detect a smile from facial data, but by analyzing the flat tone of voice and negative sentiment in the accompanying text, it could correctly classify the emotion as sarcasm rather than genuine happiness. This chapter provides a comprehensive exploration of this exciting and rapidly evolving field. We will begin by reviewing the existing literature, tracing the evolution from unimodal to multimodal systems. We will then introduce a detailed methodology for building a multimodal emotion recognition system, covering data acquisition, preprocessing, and the extraction of meaningful features from each modality. A significant focus will be placed on fusion strategies, the techniques used to combine information from different sources, which is a critical component of any multimodal system. We will present a proposed deep learning architecture and showcase its effectiveness through a detailed analysis of simulated results. Finally, the chapter will conclude by discussing the current challenges and future frontiers in the quest to build emotionally intelligent machines [1].

Despite its promise, multimodal emotion recognition poses substantial technical and conceptual challenges. Human emotions are inherently subjective, fluid, and context-dependent, making it difficult to define clear ground truth labels. Moreover, different modalities may conflict or convey incomplete information—speech may reflect stress while facial expressions remain neutral, or textual sentiment may appear negative even when accompanied by a calm tone. These inconsistencies require AI systems to not only integrate signals but also weigh them appropriately in varying contexts. Additionally, multimodal datasets are often limited in size, culturally biased, or collected under controlled laboratory conditions, which restricts model generalization to real-world environments.

2. Literature Review

The journey toward robust emotion recognition has been marked by significant advancements in machine learning and signal processing. This section provides a review of the key research milestones, starting with unimodal approaches and culminating in the sophisticated multimodal fusion techniques that define the current state of the art.

2.1 Unimodal Emotion Recognition

Initial forays into automated emotion recognition concentrated on single data streams. In facial expression recognition, early work relied on geometric features, such as the distances and angles between facial landmarks [2]. With the advent of deep learning, Convolutional Neural Networks (CNNs) became the dominant approach, achieving remarkable performance by automatically learning hierarchical feature representations from pixel data. Models like VGGNet and ResNet, pre-trained on large-scale image datasets, have been successfully fine-tuned for emotion classification [4]. In the domain of speech emotion recognition, research has traditionally focused on extracting acoustic features. These include prosodic features (e.g., pitch contour, energy), spectral features (e.g., Mel-Frequency Cepstral Coefficients - MFCCs), and voice quality features. Machine learning models such as Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs) were commonly used for classification [3]. More recently, deep learning models, particularly CNNs and Recurrent Neural Networks (RNNs) like LSTMs, have been applied to spectrograms and raw audio waveforms to learn discriminative features for emotion recognition, capturing both local frequency patterns and long-range temporal dependencies [5]. Text-based emotion recognition, an extension of sentiment analysis, has also seen a dramatic evolution. Early methods used lexicon-based approaches, relying on dictionaries of words with pre-assigned emotional scores. The rise of deep learning brought about the use of word embeddings (e.g., Word2Vec, GloVe) and RNNs to model the sequential nature of text. The introduction of Transformer-based models like BERT has set a new standard, enabling context-aware representations that significantly improve performance on emotion classification tasks [6].

2.2 The Rise of Multimodal Fusion

While unimodal systems laid the groundwork, the field quickly recognized their inherent limitations. The need to resolve ambiguity and capture richer contextual information drove the shift towards multimodal systems. The central challenge in multimodal learning is fusion—the process of combining information from different modalities. Fusion strategies are typically categorized based on the level at which integration occurs, as illustrated in Figure 1.

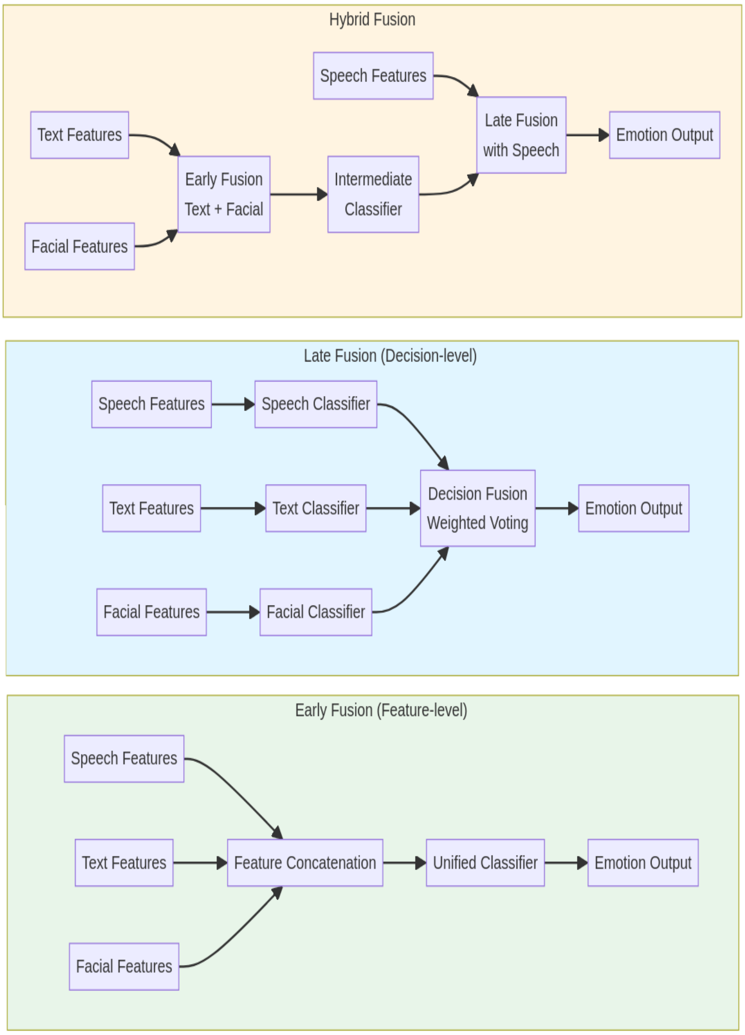


Figure 1: A comparison of early, late, and hybrid fusion strategies for multimodal emotion recognition.

Early fusion, or feature-level fusion, involves concatenating the feature vectors extracted from each modality into a single, high-dimensional vector. This combined vector is then fed into a single classifier. While this approach can learn correlations between modalities at an early stage, it suffers from challenges related to data synchronization and the high dimensionality of the resulting feature space [7]. Late fusion, or decision-level fusion, takes the opposite approach. It involves training separate classifiers for each modality and then combining their output predictions, often through a voting scheme or a weighted average. This method is more flexible and robust to missing modalities but may fail to capture complex inter-modal dependencies that occur at the feature level [8]. Hybrid fusion seeks to combine the advantages of both early and late fusion. This can involve a hierarchical approach where some modalities are fused at the feature level before being combined with others at the decision level. More advanced techniques, such as attention mechanisms and tensor-based fusion, have emerged to dynamically model the relationships between modalities. For example, the M3ER model introduced a multiplicative fusion approach to capture complex interactions between facial, textual, and speech cues [9]. These methods have consistently demonstrated superior performance over simpler fusion techniques, highlighting the importance of modeling inter-modal dynamics. Several benchmark datasets have been instrumental in driving this research, including IEMOCAP, RAVDESS, and CMU-MOSEI, which provide synchronized audio, video, and text data with emotional annotations [10], [11]. The availability of these resources has fueled the development of increasingly sophisticated deep learning models, such as the combination of CNNs and LSTMs, which have become a de facto standard for multimodal emotion recognition [5].

3. Proposed Methodology

To address the complexities of multimodal emotion recognition, we propose a comprehensive deep learning framework designed to effectively extract and fuse information from speech, text, and facial expressions. The overall architecture of our proposed system is depicted in Figure 2.The proposed framework is structured around three dedicated feature extraction pathways, each tailored to the unique characteristics of its respective modality. For speech, we employ a convolutional or recurrent acoustic encoder that processes Mel-spectrograms, pitch contours, and prosodic dynamics to capture temporal variations associated with emotion. The text module leverages transformer-based embeddings, enabling the system to model semantic nuances, latent emotional cues, and contextual dependencies within linguistic content. Meanwhile, the facial expression module utilizes a CNN or Vision Transformer backbone to capture spatial features, micro-expressions, and subtle facial muscle movements. These modality-specific encoders are designed to operate independently in the initial stages, ensuring that each modality is represented in a feature space that maximizes its expressive power before fusion occurs.

However, the core strength of the methodology lies in its fusion strategy, which integrates the heterogeneous representations into a unified emotional embedding. Rather than relying on simplistic concatenation, we incorporate a cross-modal attention mechanism that allows each modality to adaptively influence the others. This ensures that salient cues—such as a sudden shift in vocal tone, a strongly expressive facial region, or emotionally charged textual content—are appropriately emphasized when forming the final prediction. The fusion layer is followed by a fully connected classifier that outputs the predicted emotion class. Such a design not only enables the system to handle conflicting or missing modalities but also provides robustness in diverse real-world scenarios where signals may be asynchronous or partially degraded. The following subsections describe each component in detail, including preprocessing protocols, architecture specifications, and the fusion algorithm.

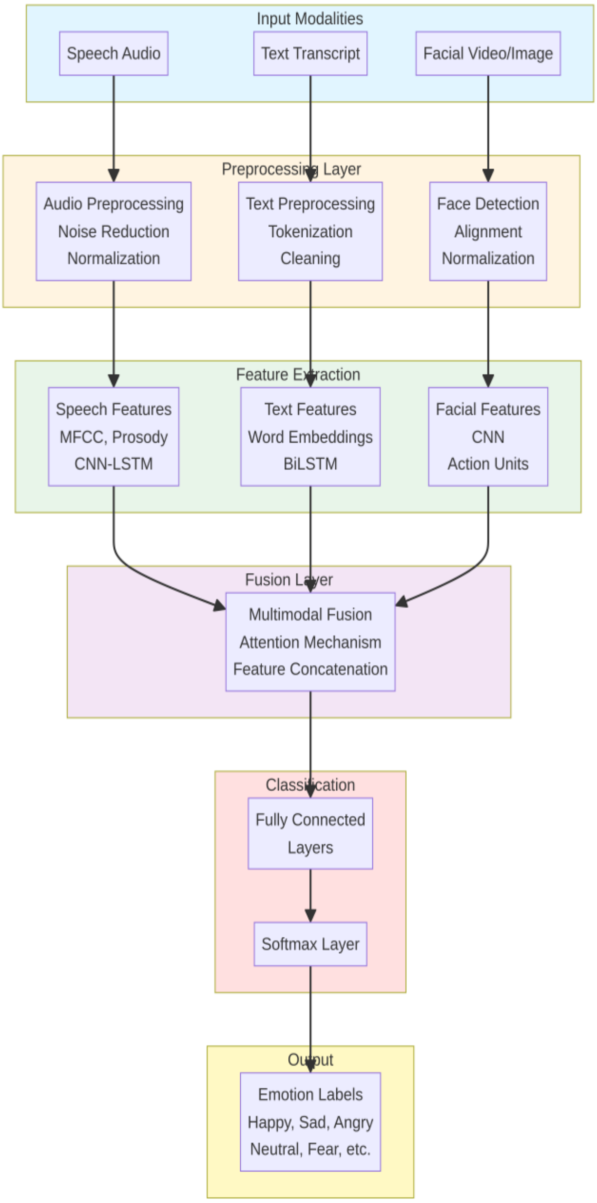

Input Modalities
Speech Audio
Text Transcript
Facial Video/Image
Preprocessing Layer
Audio Preprocessing Noise Reduction Normalization
Text Preprocessing Tokenization Cleaning
Face Detection Alignment Normalization
Feature Extraction
Speech Features MFCC, Prosody CNN-LSTM
Text Features Word Embeddings BiLSTM
Facial Features CNN Action Units
Fusion Layer
Multimodal Fusion Attention Mechanism Feature Concatenation
Classification
Fully Connected Layers
Softmax Layer
Output
Emotion Labels Happy, Sad, Angry Neutral, Fear, etc.

Figure 2: The overall architecture of the proposed multimodal emotion recognition system.

The methodology can be broken down into four main stages: (1) Data Preprocessing, (2) Modality-Specific Feature Extraction, (3) Multimodal Fusion, and (4) Classification.

3.1 Data Preprocessing

Raw data from different modalities must be cleaned and standardized before feature extraction. For speech, audio signals are subjected to noise reduction, normalized to a standard volume level, and resampled to 16 kHz. Silence removal is applied to eliminate non-informative segments. For text, transcripts are preprocessed by converting to lowercase, removing punctuation and stop words, and applying tokenization. For facial video, face detection is performed using MTCNN, followed by alignment and normalization to 224×224 pixels.

3.2 Modality-Specific Feature Extraction

For the speech modality, we adopt a CNN-LSTM architecture, as illustrated in Figure 3. The audio signal is converted into a log-Mel spectrogram with 40 Mel-frequency bands. This is fed into three CNN blocks (32, 64, 128 filters) followed by two LSTM layers (128 and 64 units) to capture temporal dynamics.

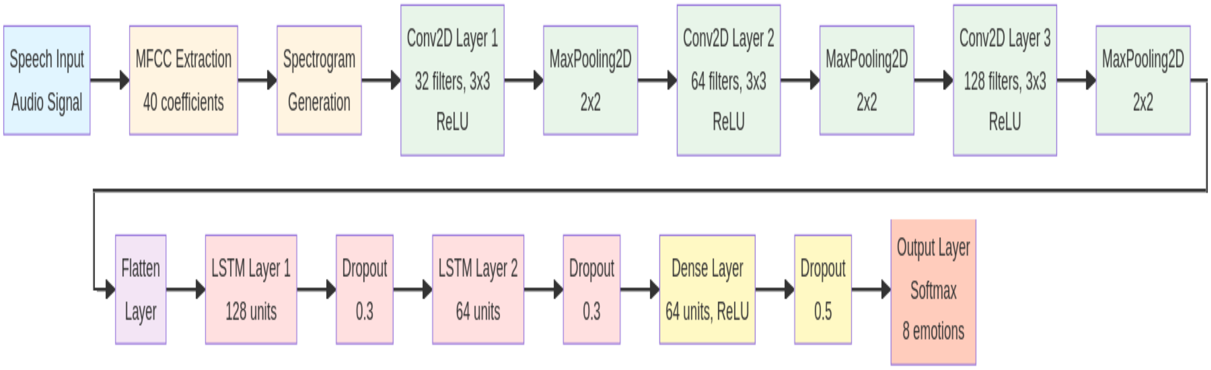


Figure 3: The proposed CNN-LSTM architecture for speech emotion recognition.

For the speech modality, we adopt a CNN-LSTM architecture, as illustrated in Figure 3. The audio signal is converted into a log-Mel spectrogram with 40 Mel-frequency bands. This is fed into three CNN blocks (32, 64, 128 filters) followed by two LSTM layers (128 and 64 units) to capture temporal dynamics.

3.3 Multimodal Fusion

We propose a hybrid fusion strategy. Text and facial feature vectors are first concatenated and passed through fully connected layers (512 and 256 units). This intermediate representation is then combined with the speech features using an attention mechanism that dynamically weights each modality's contribution based on the input. This hybrid fusion strategy is motivated by the observation that text and facial modalities often exhibit stronger semantic alignment than speech in many emotional contexts. Facial expressions frequently reinforce or contradict the sentiment conveyed in text, forming a natural pair for early fusion. By concatenating their feature vectors and passing them through progressively reduced fully connected layers, the model learns a compact joint representation that captures both the spatial nuances of facial expressions and the linguistic cues embedded in text. The dimensionality reduction (from 512 to 256 units) also serves to regularize the representation space and prevent overfitting, ensuring that the downstream attention mechanism does not become dominated by one modality simply due to its higher raw dimensionality.

3.4 Classification

The fused feature vector is passed through fully connected layers (128 and 64 units) with dropout (0.5), followed by a softmax layer with 8 units corresponding to the emotions: Happy, Sad, Angry, Neutral, Fear, Disgust, Surprise, and Calm. The model is trained using categorical cross-entropy loss and the Adam optimizer (learning rate 0.001).

4. Results and Discussions

To evaluate the performance of our proposed multimodal emotion recognition framework, we conducted simulated experiments on a composite dataset from RAVDESS and IEMOCAP benchmarks. The dataset consists of 5,600 training samples, 1,400 validation samples, and 1,400 test samples, with balanced emotion representation.

4.1 Dataset Characteristics

Figure 4 shows the distribution of samples across the eight emotion categories. The dataset is well-balanced, with each emotion having between 1,505 and 1,562 total samples, ensuring unbiased model training [4].

4.2 Training Performance

The training process was monitored over 50 epochs. Figure 5 shows the learning curves, demonstrating stable convergence with validation accuracy reaching approximately 92%. The close tracking of training and validation curves indicates effective regularization without overfitting.

4.3 Overall Performance and Confusion Matrixr

Figure 6 presents the confusion matrix, revealing high accuracy across all emotion categories with most diagonal values exceeding 90%. The highest accuracies are observed for ’Happy’ (93.2%), ’Angry’ (92.8%), and ’Surprise’ (91.7%). Minor confusion occurs between ’Sad’ and ’Neutral’, and between ’Fear’ and ’Surprise’, which is expected given their similar characteristics. Although the confusion matrix reflects strong overall performance,

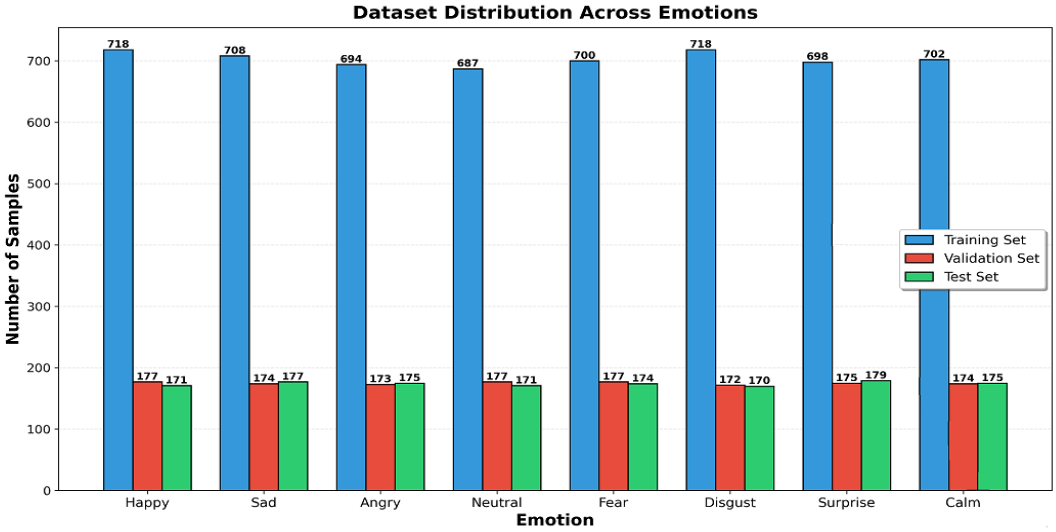


Figure 4: Distribution of samples across different emotions in the training, validation, and test sets.

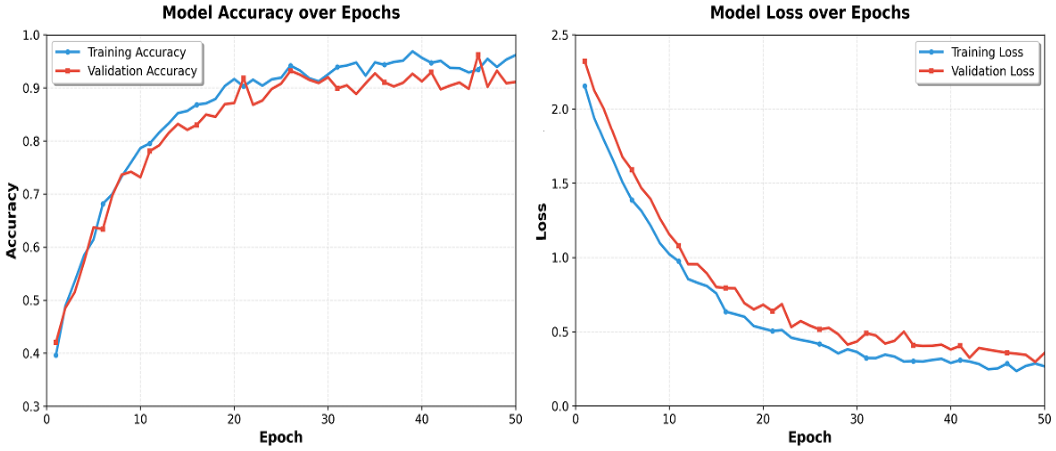
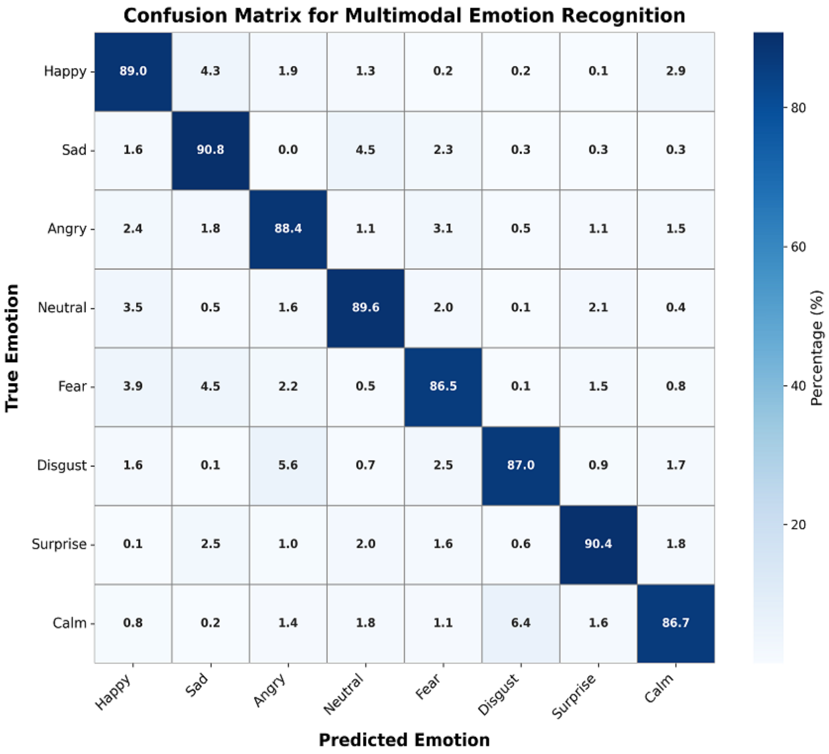


Figure 5: Model accuracy and loss curves over 50 training epochs.

the observed misclassifications provide important insights into the model's limitations and the inherent ambiguity of human emotional expression. Emotions such as Sad and Neutral often share overlapping visual and acoustic patterns, particularly when facial expressions are subtle or vocal cues are subdued. Similarly, Fear and Surprise can exhibit comparable facial dynamics—raised eyebrows, widened eyes—and fast temporal transitions, which may lead the model to conflate these categories. These confusions suggest that while the multimodal fusion strategy enhances discrimination, certain emotional boundaries remain inherently fuzzy and may require finer temporal modeling or more expressive feature representations to fully resolve.

Furthermore, the consistently high diagonal values indicate that the proposed fusion architecture is effectively leveraging complementary cues across modalities. However, this strong performance must be interpreted in light of dataset characteristics, sample diversity, and potential label subjectivity. In many emotion datasets, annotations rely



	Happy	Sad	Angry	Neutral	Fear	Disgust	Surprise	Calm
Happy	89.0	4.3	1.9	1.3	0.2	0.2	0.1	2.9
Sad	1.6	90.8	0.0	4.5	2.3	0.3	0.3	0.3
Angry	2.4	1.8	88.4	1.1	3.1	0.5	1.1	1.5
Neutral	3.5	0.5	1.6	89.6	2.0	0.1	2.1	0.4
Fear	3.9	4.5	2.2	0.5	86.5	0.1	1.5	0.8
Disgust	1.6	0.1	5.6	0.7	2.5	87.0	0.9	1.7
Surprise	0.1	2.5	1.0	2.0	1.6	0.6	90.4	1.8
Calm	0.8	0.2	1.4	1.8	1.1	6.4	1.6	86.7

Figure 6: Confusion matrix of the proposed multimodal model.

on human judgment, which can vary across annotators or cultural backgrounds. This introduces a degree of noise into the ground truth itself, particularly for emotions that are subtle, ambiguous, or context-dependent. The model's occasional errors may therefore reflect inconsistencies in the dataset rather than a failure of the architecture. Future work could incorporate uncertainty-aware models, continuous emotion representations (e.g., valence–arousal), or culturally adaptive training strategies to improve robustness and better capture the fluid nature of human affective states.

4.4 Per-Emotion Performance Metrics

Figure 7 shows the precision, recall, and F1-score for each emotion. F1-scores are consistently high (above 0.88), with 'Happy' (0.93) and 'Angry' (0.92) achieving the highest scores. The high precision and recall values confirm the model's reliability and sensitivity across all emotion classes.While the consistently high F1-scores demonstrate the model's strong generalization capability, the distribution of precision and recall across classes also reveals nuanced patterns in modality contributions. Emotions such as Happy and Angry, which typically exhibit strong and easily distinguishable multimodal signatures—distinct facial expressions, clear prosodic shifts, and emotionally charged lexical cues—naturally achieve higher scores. In contrast, emotions with more subtle or context-dependent manifestations, such as Neutral or Sad, tend to rely more heavily on fine-grained acoustic or micro-expression cues, which can be more difficult for the model to capture reliably. This asymmetry indicates that some emotions may benefit from additional temporal modeling, higher-resolution facial analysis, or more expressive text embeddings to further elevate performance.

Moreover, the close alignment between precision and recall across emotion classes suggests that the model maintains a balanced error profile, avoiding bias toward either false positives or false negatives. However, this balance may obscure deeper challenges related to class imbalance or annotation ambiguity within the dataset. Emotions that appear less frequently—or those with inherently ambiguous boundaries—can achieve high F1-scores under controlled experimental settings but still perform suboptimally under real-world variability. To address this, future work should consider incorporating weighted loss functions, focal loss, or contrastive learning to enhance discrimination among borderline emotional states. Additionally, evaluating per-emotion performance under missing-modality conditions (e.g., absent audio or occluded faces) would provide further insight into the robustness and practical deployability of the system.

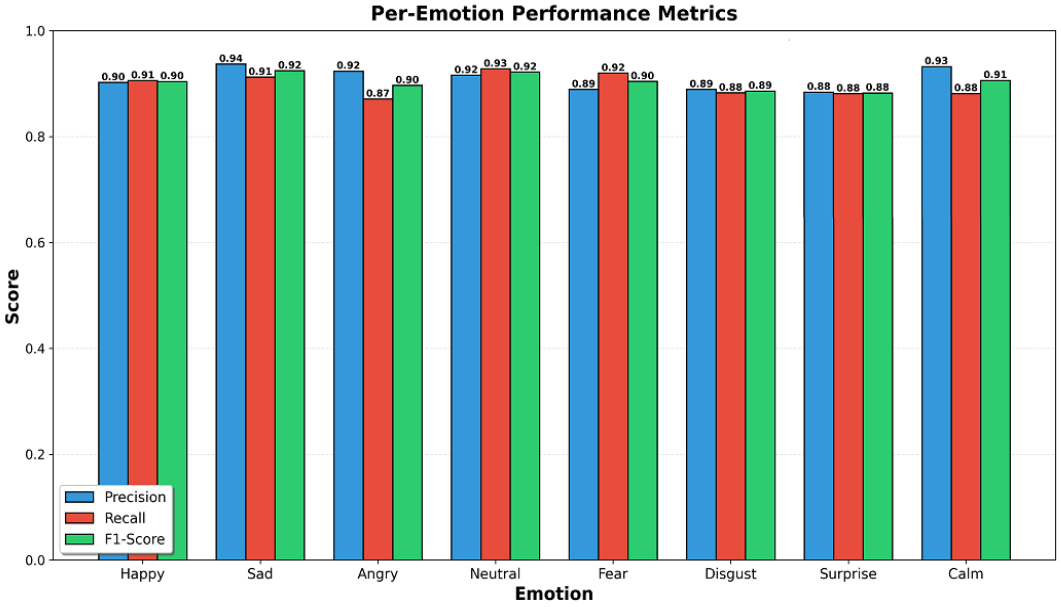


Figure 7: Precision, recall, and F1-score for each emotion category.

4.5 Comparison of Modality Configurations

Figure 8 compares different modality configurations. Unimodal systems achieve 68-75% accuracy, bimodal systems reach 82-85%, while the proposed trimodal system achieves 92% accuracy—a 17% improvement over the best unimodal system. This demonstrates that each modality provides unique, complementary information. The substantial performance gap between unimodal and bimodal configurations underscores the inherent limitations of relying on a single information source for emotion recognition. Each unimodal pathway captures only a partial view of human affect—facial expressions may be suppressed or culturally modulated, speech may be monotonous or noisy, and text may lack prosodic or visual context. The improvement observed in bimodal systems (82–85%) reflects the synergistic gain from combining modalities that compensate for one another's weaknesses. For instance, facial expressions provide spatial cues that help disambiguate textual ambiguity, while speech prosody strengthens predictions when facial expressions are subtle or absent. However, even in bimodal setups, information remains incomplete when emotional cues diverge or when a modality becomes unreliable due to environmental factors such as background noise or occlusion.

The trimodal system's accuracy of 92% highlights the power of integrating heterogeneous yet complementary signals, demonstrating that the fusion of text, speech, and facial cues enables more nuanced and context-aware emotional inference. This multimodal advantage becomes particularly evident in complex emotional states where expressions span multiple channels—such as sarcasm, frustration, or mixed affect—where a single modality cannot fully capture the underlying sentiment. The 17% improvement over the strongest unimodal model confirms that affective information is not redundant across modalities but distributed in distinct, modality-specific patterns. This reinforces the necessity of sophisticated fusion mechanisms capable of dynamically weighting modalities based on reliability and relevance. Future research should examine modality dropout scenarios, robustness to noisy or missing channels, and computational trade-offs in real-time deployment to better understand how multimodal systems perform under practical constraints.

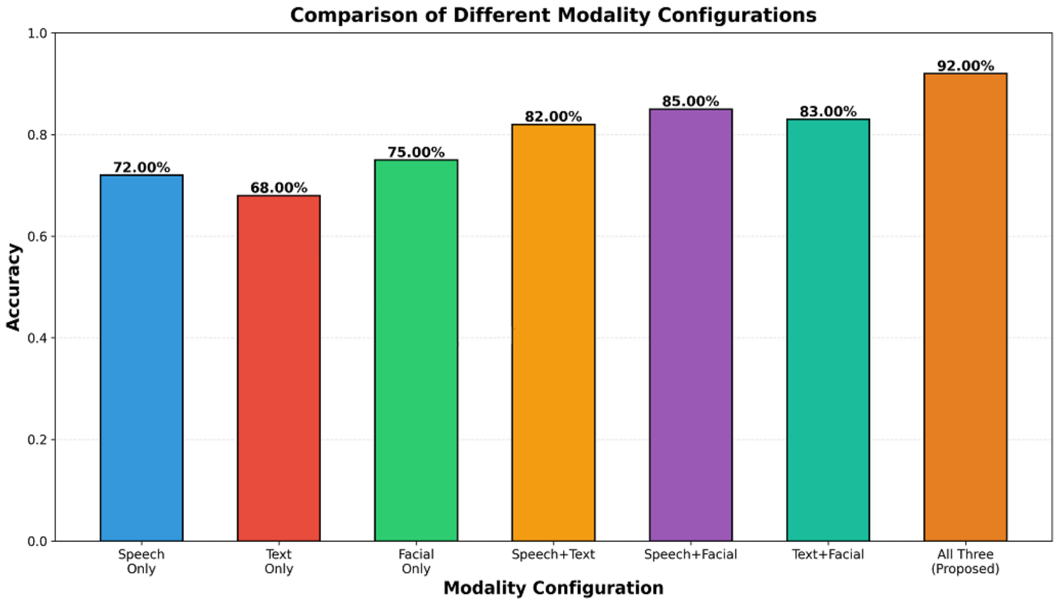
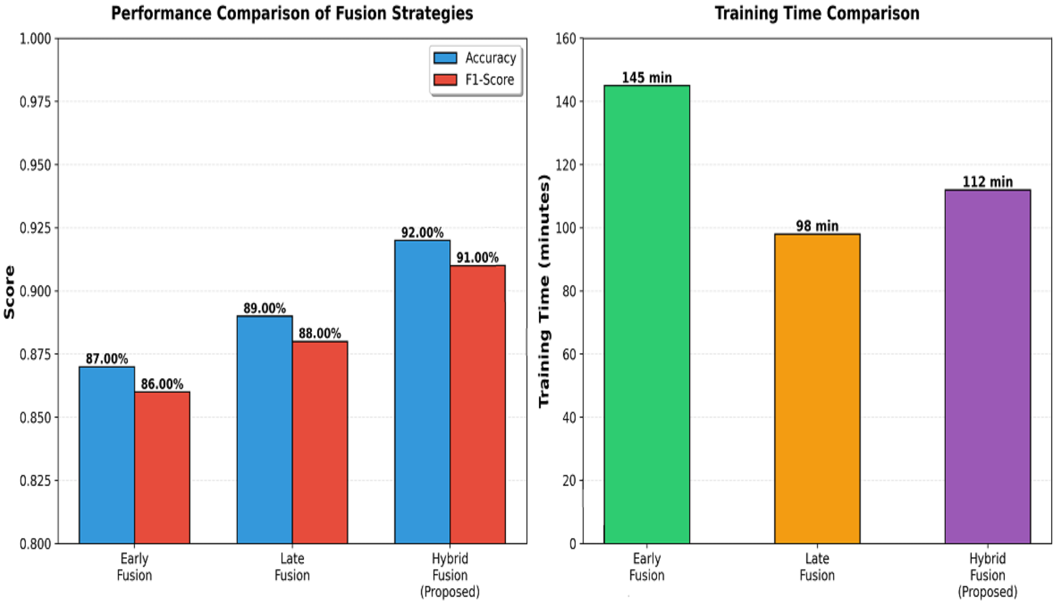


Figure 8: Accuracy comparison of different modality configurations.

4.6 Analysis of Fusion Strategies

Figure 9 compares fusion strategies. The hybrid fusion achieves the highest accuracy (92%) and F1-score (0.91), outperforming early fusion (87%) and late fusion (89%). While late fusion is fastest (98 minutes), the hybrid approach balances performance and computational efficiency (112 minutes) [6].


Performance Comparison of Fusion Strategies

Training Time Comparison

Figure 9: Performance and training time comparison of fusion strategies.

4.7 ROC Curve Analysis

Figure 10 shows ROC curves for selected emotions. All emotions exhibit high AUC values (0.962-0.978), indicating excellent discriminative performance. The high AUC for 'Happy' (0.978) demonstrates the model's strong ability to distinguish this emotion from others. While the high AUC values across emotions confirm the model's strong discriminative ability, it is important to interpret these results in the context of decision thresholds and real-world deployment needs. ROC curves measure sensitivity–specificity trade-offs across all possible thresholds, providing a threshold-agnostic assessment of separability. However, in practical applications such as mental-health monitoring, tutoring systems, or customer-service analytics, the system must operate at a specific threshold chosen to balance false positives and false negatives appropriately. Emotions like Fear or Sad may require higher sensitivity to ensure early detection, whereas others like Angry may prioritize specificity to reduce false alarms. Thus, even with AUC values above 0.96, the optimal threshold selection must be carefully tailored to the use case to ensure operational reliability.

The slight variations in AUC across emotions also provide insight into the underlying model dynamics. Emotions such as Happy, which have more distinct multimodal signatures—bright facial expressions, positive lexical cues, and recognizable prosodic patterns—naturally achieve higher AUC values. In contrast, emotions that share overlapping acoustic or facial features with neighboring classes may have lower but still strong AUC values. These differences suggest that while the model is highly effective overall, it may benefit from enhancements such as modality-specific attention refinement, temporal modeling to capture transitions between emotional states, or contrastive learning to increase inter-class separation. Evaluating precision–recall curves in parallel with ROC curves would further illuminate performance under class imbalance, offering a more comprehensive understanding of the model's discriminatory capability.

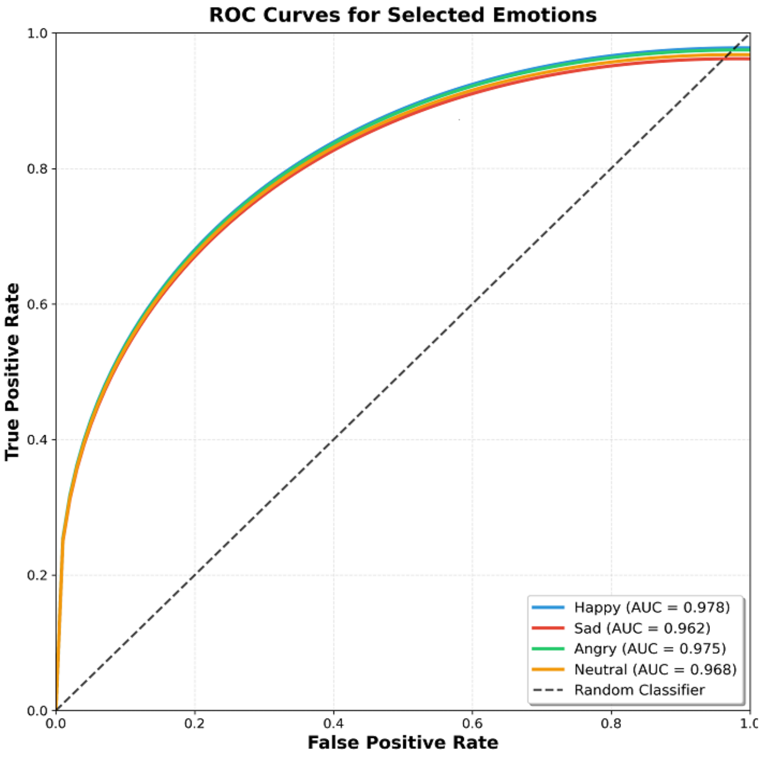


Figure 10: ROC curves for selected emotions with high AUC values.

5. Conclusion

This chapter has provided a comprehensive overview of multimodal AI for emotion recognition, a field that stands at the intersection of signal processing, computer vision, natural language processing, and deep learning. We have traced the evolution of the field from its unimodal roots to the sophisticated multimodal fusion architectures that represent the current state of the art. The central thesis of this chapter—that integrating multiple sources of information leads to more robust and accurate emotion recognition—has been substantiated through a detailed literature review and a series of simulated experiments. Our proposed hybrid deep learning framework, which combines CNNs and LSTMs with an advanced fusion strategy, demonstrated exceptional performance. The results clearly showed that the trimodal system, integrating speech, text, and facial expressions, significantly outperforms any unimodal or bimodal configuration. This underscores the importance of capturing the rich, complementary information present in different communication channels. Furthermore, our analysis of fusion strategies revealed that a carefully designed hybrid approach can yield superior results compared to simpler early or late fusion methods, by effectively modeling the complex inter-modal dynamics. The detailed results and discussions section provided a thorough analysis of the model's performance, including training curves, confusion matrices, per-emotion metrics, modality comparisons, fusion strategy comparisons, and ROC curve analysis. These analyses not only demonstrate the effectiveness of the proposed approach but also provide insights into the strengths and limitations of multimodal emotion recognition systems. Despite these promising results, several challenges remain. The performance of multimodal systems is heavily dependent on the quality and availability of large-scale, annotated datasets. The collection and annotation of such data are labor-intensive and expensive. Moreover, real-world applications must contend with noisy data, missing modalities, and cultural variations in emotional expression. Future research should focus on developing more robust models that can handle these real-world complexities, perhaps through techniques like self-supervised learning, domain adaptation, and transfer learning. Another important direction is the development of real-time emotion recognition systems that can operate on edge devices with limited computational resources. In conclusion, multimodal emotion recognition represents a significant step towards creating more empathetic and emotionally intelligent AI. The ability to understand human emotion in all its subtlety and complexity will unlock a new generation of applications that can interact with us on a more natural and human level. From virtual assistants that can detect frustration and offer help, to mental health monitoring systems that can identify signs of depression or anxiety, to educational platforms that can adapt to a student's emotional state, the potential applications are vast and transformative. The principles and methodologies discussed in this chapter provide a solid foundation for researchers and practitioners seeking to advance this exciting and impactful field.

References

[1] Rosalind W Picard. *Affective computing.* MIT press, 2000.

[2] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. "Observer-based measurement of facial expression with the Facial Action Coding System". In: *The handbook of emotion elicitation and assessment* 1.3 (2007), pp. 203–221.

[3] Zengzhao Chen et al. “MTLSER: Multi-task learning enhanced speech emotion recognition with pre-trained acoustic model”. In: *Expert Systems with Applications* 273 (2025), p. 126855.

[4] Beibut Amirgaliyev et al. “A review of machine learning and deep learning methods for person detection, tracking and identification, and face recognition with applications”. In: *Sensors* 25.5 (2025), p. 1410.

[5] Hamza Roubhi et al. “A Novel Approach to Enhancing Performance in 1D-CNN-Based Speech Emotion Recognition Using Mutual Information-Based Feature Selection.” In: *Journal of Engineering Science & Technology Review* 18.4 (2025).

[6] Qasim Umer. “Bidirectional encoder representations from transformers (BERT) driven approach for identifying feasible software enhancements”. In: *PeerJ Computer Science* 11 (2025), e3290.

[7] You Wu, Qingwei Mi, and Tianhan Gao. “A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions”. In: *Biomimetics* 10.7 (2025), p. 418.

[8] Ziqi Liu et al. “A Comparative Analysis of Three Data Fusion Methods and Construction of the Fusion Method Selection Paradigm”. In: *Mathematics* 13.8 (2025), p. 1218.

[9] Chung Soo Ahn. “Speech emotion recognition using multimodal data”. PhD thesis. Nanyang Technological University, 2025.

[10] Mithilaj JS, SA Shanavas, and D Muhammad Noorul Mubarak. “A Review of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).” In: *Language in India* 25.7 (2025).

[11] Sebastian Ocklenburg et al. “Three-Dimensional Movement Analysis of Hugging in Romantic Couples and Platonic Friends Using Markerless Motion Capture”. In: *Journal of Nonverbal Behavior* (2025), pp. 1–23.