iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83881

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Multimodal Emotion Recognition: A Comprehensive Survey of Architectures, Fusion Strategies, Datasets, and Future Directions

Faisal Majeed[1], Poonam Dhankhar[2]
Department of Computer Science and Engineering, Ganga Institute of Technology and Management, Kablana, India

Abstract: ***Recent advancements in computer science is showing numerous miracles one reason is human recognition framework which is considered to be the base for human computer Interaction (HCI). This functionality reduces the gap between artificial empathic & socially aware systems. In this regard various developments have been made. The early models were built by keeping in view the single factor for recognizing the human emotions which include the facial expressions, voice tone, facial gestures, eye movements etc. In reality the human emotions are properly recognized when we consider the mentioned multiple factors into consideration at once. These features range from what a person says, how their voice changes pitch, face gestures various other physiological signals such as heart rate or skin responses. All these features collectively allow any system to recognize human emotions accurately. Unimodal emotion detection systems process only a single type of modality at a time and often fail to capture complex emotional states, but Multimodal system removes this problem by combining all the features and collectively give a result on the basis of various feature all at once and has shown remarkable results. This survey provides an overview and a deeper understanding of the state-of-the-art Multimodal Emotion Recognition systems. The paper starts with analyzing classical methods to the recent multimodal emotion detection systems that are predominantly based on Transformers architectures. They leverage pretrained models like Vision Transformer on facial features, Wav2Vec 2.0 on speech and BERT for text. These features are then fused via cross-attention or multimodal transformers. High-end systems might leverage the latest large multimodal models that can take images, audio and text together. In addition, modern multimodal emotion recognition systems rely on CNNs, LSTMs, CNN-LSTM hybrids, graph neural networks, autoencoders, capsule networks and ensemble methods. Older systems relied on traditional CNN and LSTM and newer systems are using more graph-based approaches as well as large multimodal foundation models.***

Keywords: Multimodal Emotion Recognition, Affective Computing, Cross-Modal Fusion, Multimodal Large Language Models, Contrastive Learning, Embodied AI, Graph Neural Networks.

I. INTRODUCTION

Emotion recognition refers to detecting the emotional state of a person based on behavioral and physiological signals including facial expressions, speech, text and body signals. [1] By equipping computers with the ability to recognize emotion, these systems have revolutionized interaction between humans and machines. When you speak, modern systems understand the emotional state of a user and respond in a manner appropriate to that state of physiology both intelligently and contextually. This enables a more medium-like way for computers to act in service of the user, in a more partnering and supportive role, rather than as passive tools. [3] Correctly identifying human emotions paves the way for huge potential uses across many domains of usage. In healthcare and psychiatry, emotion recognition systems can monitor patients suffering from mental disorders such as depression, anxiety, and neurological disabilities. Tracking a patient's emotional state over time will provide valuable insights which these systems will then convert into objective biometric data that homes in on the individual and track each day after treatment with their doctors and therapists. Emotion-aware learning systems can identify the learning deficits of students [5], for example when students demonstrate frustration, confusion, or lack any interest in their studies. And according to their behavior the changes can be brought in their learning rate, content preferences and adding more assistance depending on these emotional signals. This provides a better outcome in their learning experience which keeps the students overall motivated and engaged. [2] Moreover, in the industries like social robotics, driver monitoring systems and virtual assistants, emotion recognition systems are quite crucial to understand the people's emotions.

©IJRASET: All Rights are Reserved

In the past, most of the Emotion Recognition systems were using only one type of information such as facial expressions or speech signals. They are named as unimodal emotion recognition systems. These systems relied on handcrafted features and did a very good job in controlled settings but frequently were failed to capture the emotions of a person properly. These features were extracted using algorithms like Local Binary Patterns (LBP) were utilized to capture facial patterns, Mel-Frequency Cepstral Coefficients (MFCCs) extracted features of speech signals. These features were then fed into traditional Machine learning algorithms for the classification of emotions, like SVM was used for classification [1]. The unimodal emotion recognition systems using one modality are very vulnerable and quite unstable, because human emotions are not implied properly using single channel, but depend on multiple data inputs a time [10]. For example, a loud noise sometimes means high level of engagement or may represent high levels of anger. So, the emotion recognition requires additional information which may be in the form of facial expressions or the spoken words or any context related information [2]. This problem of insufficiency is solved by using a model, known as Multimodal Emotion Recognition (MER). MER model combines information from different inputs such as facial expressions, speech, text etc. and combine them using fusion models and outputs the emotion details of a person. The reason these techniques yield superior performance is that they capitalize on two key advantages: redundancy and complementarity. Redundancy increases the reliability of the system since if one sensor is absent or noisy, the other sensors can still yield information. Complementarity which helps in Diversifying the training wherein each modal provides various intricacies to enable emotion recognition and gives improvement in accuracy. [12] The development of Multimodal Emotion Recognition (MER) has progressed rapidly with the introduction of deep learning. [14] In these systems, deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are used to automatically extract important spatial and temporal patterns directly from raw data and this leads to more accurate and effective recognition of emotions. Apart from these models the Transformer models are becoming more popular in the MER systems. The Transformer models are very much effective to capture the long-range interactions and relations between various data modalities of text, speech and facial expressions. The modern Multimodal Large Language Models (MLLMs) have given a new direction to the emotion recognition systems. These models not only categories the emotions but also provide the explanations about the predicted emotions [15]. The Multimodal Emotion Recognition (MER) has made significant progress in predicting correct outputs but its research field is still fragmented. There still exists many important challenges. One of the major issues in MER systems is that while dealing with different modalities, such as speech, text and video, the systems are not able to align perfectly in time. Another challenge includes that a model performance drops sharply when one or more modalities are missing. The other type of issue is related to the datasets. The datasets suffer from class imbalance i.e., some emotions are present more frequently, while others are not, this effects the overall system's learning and performance. The previous surveys mainly have focused on the specific areas of text-based analysis and body gestural signal processing. This survey brings a detailed review of modern deep learning models and fusion models. The overall architecture of MER is explained in the *Figure 1.*

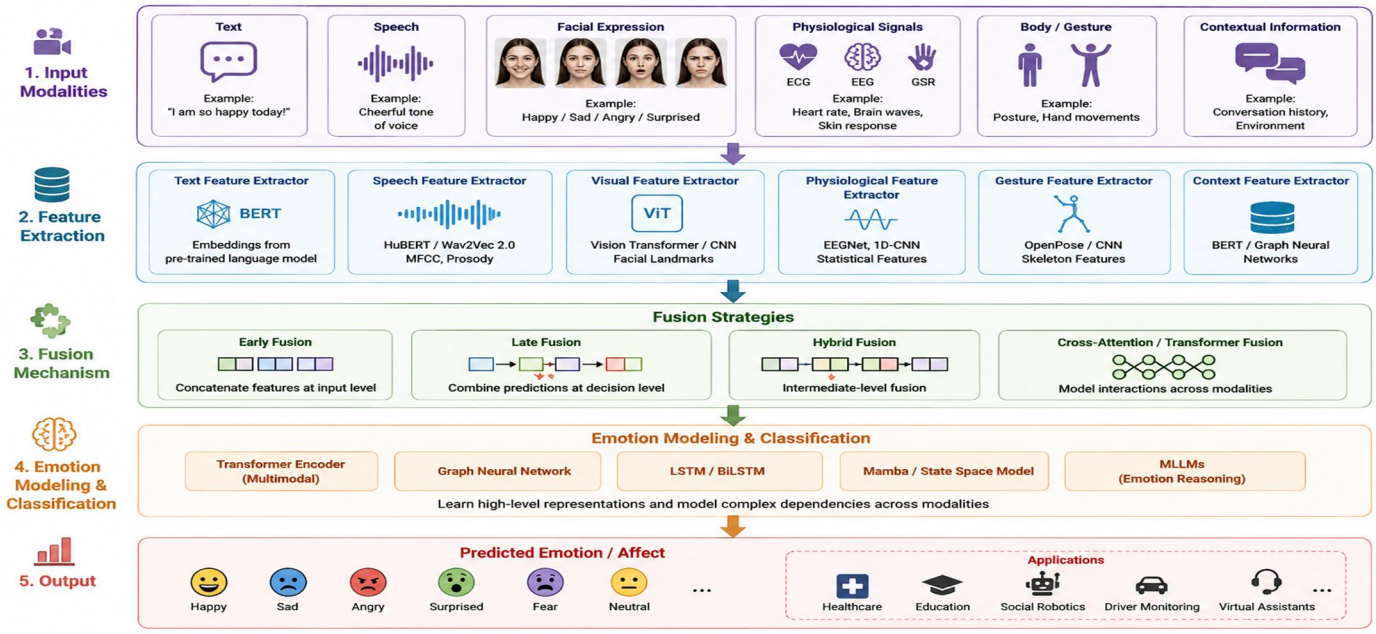


Figure 1: Overall architecture of MER systems

II. MAIN POINTS OF CONTRIBUTION OF THIS SURVEY

1) The paper provides a detailed analysis of feature learning from text, audio, visual, physiological, and contextual data for emotion recognition.
2) Critically reviewed the evolution of multimodal emotion recognition architectures, comparing CNNs, RNNs, Graph Neural Networks, State Space Models, and Multimodal Large Language Models (MLLMs). The paper discusses these models on the basis of computational cost, performance and practical suitability.
3) A systemic mathematical framework is developed for multimodal fusion algorithms which also includes cross-modal attention mechanisms and self-supervised learning techniques.
4) The paper also reviews major benchmark datasets such as IEMOCAP, MELD, CMU-MOSEI and DEAP and highlights their key limitations which include inconsistent annotations and biases based on demography specially towards WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations.
5) It also gives the future research directions. Using the federated learning for privacy protection and edge AI for real time output processing.

III. BACKGROUND AND FUNDAMENTALS

A. *Affective Computing*

Affective computing is a multi-domain concept which combines concepts from various subjects such as computer science, psychology and cognitive neuroscience. It is defined as branch of artificial intelligence that focuses of studying and developing of a system that can recognize, interpret, process, and simulate human emotions. The development of such systems offers wide range of challenges and the most inherent challenge lies in bridging the semantic gap between low-level sensory inputs of computer (e.g., pixel intensities, audio patterns, voltage changes) and high-level abstract psychological meanings which humans can understand. [4]

B. *Emotion Models*

In machine learning based emotion recognition, emotions need to be defined in a measurable form. Therefore, most MER studies use two main psychological models to label and classify emotions.

1) Ekman's Basic Emotions (Categorical Models): On the basis of psychological research of Paul Ekman, the categorical emotion model suggests that human emotions are categorised into following discrete classes: Anger, Disgust, Fear, Happiness, Sadness, and Surprise, frequently augmented with a Neutral baseline [7]
   This model makes emotion recognition as a classification problem in machine learning. The classification problems are usually dealt with cross-entropy loss functions. But in reality, the human emotions are far more complex to be classified using a simple classification algorithm. Because emotion states of humans overlap with one another, and evolve with time, or can occur simultaneously as mixed emotions. Therefore, the categorical models may fail to capture the emotional variations and changes properly. [9]
2) Dimensional Models (Valence-Arousal-Dominance): The problem of discrete emotion categories in categorical models was removed by introduction of dimensional emotion models. The James Russell's "Circumplex Model of Affect" is most widely used approach. In this model, the human emotions are represented using a continuous multidimensional space, such as *valence* and *arousal* to define the emotional states [7].

- *Valence* quantifies the intrinsic hedonic tone, ranging from highly negative (unpleasant) to highly positive (pleasant).
- *Arousal* measures the degree of physiological and psychological activation, ranging from passive (calm, lethargic) to active (excited, tense).
- *Dominance* (less frequently utilized) assesses the degree of control or submissiveness the subject experiences regarding the stimulus [22].

Therefore, dimensional model makes MER a continuous regression problem instead of a fixed classification task. This approach is very useful in conversational and physiological emotion recognition, where emotions do not change abruptly but are recognised gradually with time [7]

C. *Multimodal Learning Fundamentals and Cross-Modal Representation*

Multimodal learning involves constructing computational architectures capable of processing and relating information from multiple modalities.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Let the input space be defined by a set of modalities $M = \{m_1, m_2, \dots, m_k\}$. For each modality $m \in M$, there exists a high-dimensional input sequence $X^{(m)} \in R^{\{T_m \times d_m\}}$,

where T_m represents the modality-specific temporal sequence length and d_m represents the feature dimensionality.

The primary objective of cross-modal representation learning is to project these heterogeneous inputs into a shared, joint latent manifold Z such that the semantic similarities and complementary interactions are preserved [24]. The mapping functions $f^{(m)}: X^{(m)} \rightarrow Z^{(m)}$ must handle inherently asynchronous data rates, for instance, audio sampled at 16 kHz, video at 30 frames per second, and text segmented into sub-word tokens. [1]

D. Temporal Sequence Modelling

Emotions possess significant temporal inertia; a localized expression (e.g., a fleeting micro-expression) is fundamentally conditioned by the preceding affective context [9]. Consequently, AER requires sophisticated sequence modelling. Mathematically, given a target emotion $\boldsymbol{y_t}$ at time step $\boldsymbol{t}$, the system models the conditional probability distribution based on the historical trajectory of all fused modalities up to that point:

$$\boldsymbol{P\left(y_t \mid x_{1:t}^{(m_1)}, x_{1:t}^{(m_1)}, \dots, x_{1:t}^{(m_1)}; \theta\right)}$$

Where $\boldsymbol{\theta}$ represents the learnable parameters of the deep learning architecture [24]. Modern architectures dynamically re-weight the influence of the historical context $\boldsymbol{x_{1:t-1}}$ to predict $\boldsymbol{y_t}$, filtering out transient noise and focusing on causally significant emotional triggers [27].

IV. MODALITIES USED IN EMOTION RECOGNITION

The performance of a MER system depends on the capability of feature extraction from different data channels or modals. The various modalities used on the MER system include text, speech and facial expressions. It also includes various other characteristics that can influence emotion recognition such as various time-based patterns or noise profiles.

A. Text-based Emotion Recognition

The textual data is derived from various sources such as direct text inputs, social media posts or Automatic Speech Recognition (ASR) system transcripts [9]. In Emotion Recognition of Conversation (ERC) system, the speech is considered to be the dominant modality as it directly shows the speaker's intent and meaning [14].

To extract information related to human emotions from textual data, Natural Language Processing (NLP) techniques and embedding models are used. These approaches help machine learning systems understand the various linguistic patterns, contextual meaning and emotion related information present in the text.

1) NLP Approaches and Embeddings:

The NLP algorithms are used to extract the emotion related information from the text. Earlier NLP algorithms worked mainly bu using the frequency-based techniques, such as TF-IDF and emotion lexicons. These methods were simple and used less computation power but often failed to capture the context-based meaning, idiomatic expression and negation statements. The field of NLP made significant improvements with the help of contextual embedding models, particularly with Bidirectional Encoder Representations from Transformers (BERT) and its improved models (such as RoBERTa and DeBERTa) [30]. These models generated contextual embeddings by keeping in view the preceding and following words in a sentence. This made them to better extract the semantics and underlying emotional information within text.

The newer models based on generative Large Language Models (LLMs) such as GPT-4 further improved the text-based emotion recognition by using zero-shot and context related understandings of complex emotional expressions even in ambiguous scenarios.

B. Speech and Audio Emotion Recognition

Speech Emotion Recognition (SER) works on the principle of analysing the variations in audio signals, such as tone, pitch, and intensity in order to identify emotions regardless of the actual spoken words [1]. The various acoustic and prosodic features are extracted from the speech that show the changes in vocal expressions. These features then help the machine learning models to capture the emotional patterns.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

1) *Acoustic and Prosodic Features:* The sound signals were traditionally represented by using hand-crafted low-level descriptors (LLDs). But the new features like Mel-Frequency Cepstral Coefficients (MFCCs) are quite common, as they compress the audio waveform in such a way that it closely represents human auditory perception. Apart from these spectral features, the prosodic features like pitch (or F_0**), speech** energy, speaking rate, jitter etc are important indicators of emotional state [1]. For example, emotion states such as anger or panic increase the vocal cord tension resulting in higher pitch and jitter
2) *Spectrograms and Self-Supervised Models:* In order to further improve the feature representation, audio signals are converted into time frequency visual diagrams (2D graphs) using a mathematical technique called Short-Time Fourier Transform (STFT). he audio signals are converted into 2D spectrograms or scalograms. These images are then fed into 2D-CNNs algorithms (e.g., Inception-ResNet) [12]. These algorithms are replaced by new models of self-supervised models such as Wav2Vec 2.0 and HuBERT (Hidden Unit BERT). These are Transformer based models which are pre-trained on large amounts of unlabelled speech data and work very well for different languages and noisy environments. To leverage the powerful pattern recognition capabilities of computer vision, 1D audio signals are frequently transformed into 2D time-frequency representations via the Short-Time Fourier Transform (STFT), yielding spectrograms or continuous wavelet transform (CWT) scalograms. These images are then processed using standard 2D-CNNs (e.g., Inception-ResNet) [12]. More recently, self-supervised acoustic models such as Wav2Vec 2.0 and HuBERT (Hidden Unit BERT) have revolutionized SER. Pre-trained on thousands of hours of unannotated speech, these transformer-based architectures learn deep, context-aware acoustic embeddings that drastically outperform traditional MFCCs, providing robust generalization across highly variable acoustic environments and cross-lingual datasets [34].

C. *Facial and Visual Emotion Recognition*

The facial expressions are extracted from the visual modality extracted from a video or from the continuous snapshots and is very essential for non-verbal affective decoding.

1) *CNNs, ViTs, and Facial Landmarks:* Standard visual pipelines extract frame-level spatial features using deep CNNs (e.g., ResNet, VGG) pretrained on massive image datasets.[9] To handle the temporal dimension of video sequences, 3D-CNNs or CNN-LSTM hybrids are deployed to track spatial deformations over time. Increasingly, Vision Transformers (ViTs) are supplanting CNNs; by dividing the facial image into patches and applying global self-attention, ViTs more effectively capture the long-range spatial interdependencies of facial features [9]. Alternatively, geometric approaches rely on extracting facial landmarks. The Facial Action Coding System (FACS) maps specific facial muscle contractions into discrete Action Units (AUs). Analysing the geometric displacement of these AUs provides an identity-invariant, explainable representation of facial affect, though it remains vulnerable to severe head pose variations and occlusions. [1]
2) *Micro-Expression Recognition:* The facial expressions or the macro-expressions are under the control of humans but micro-expressions are involuntary and last for very short span (usually last for between $\mathbf{1/25}$ and $\mathbf{1/5}$ of a second). These expressions appear when a person tries to conceal the true emotions consciously or unconsciously [37]. Recognition of these micro expressions are very important for an efficient MER and often come up with highly challenging task of gathering and processing the visual channel over high frame-rate. This data is then passed into specialized spatiotemporal deep learning models, such as optical flow techniques and 3D Convolutional Neural Networks (3D-CNNs) which are capable to identify and classify the small facial changes into emotions accurately [2]

D. *Physiological Signal-based Emotion Recognition*

The physiological signals are those signals which are generated by the autonomic nervous system and cannot be controlled or concealed like that of audio and visual modalities. Therefore, collective details regarding the physiological motions provides more objective indication of human emotions [4].

The physiological signals are collected by the following techniques:

1) *Electroencephalography (EEG)***:** It is a technique to record the electrical activity of the human brain with very high temporal resolution [4]. The signals recorded by EEG for emotion recognition is divided into different frequency bands. The signals recorded are delta, theta, alpha, beta and gamma. They are recorded by using techniques such as Fourier Transform or Wavelet Transform. These frequency bands are connected with different emotional and cognitive states. For example, beta and gamma waves are often linked to increased mental activity and emotional arousal, while frontal alpha asymmetry has been widely studied as an indicator of emotional valence which reflects the human's approach behaviour and withdrawal behaviours [41, 42].

2) *ECG (Electrocardiography) & GSR (Galvanic Skin Response)*: Peripheral physiological signals are those signals which provide information about the activation of the sympathetic nervous system. This information is closely related to emotional and stress responses. In emotion recognition, Electrocardiography (ECG) is commonly used to measure Heart Rate Variability (HRV), in which both the time and frequency-based features are analysed simultaneously to detect emotional behaviour and stress levels [4]. Similarly, Galvanic Skin Response (GSR), also known as Electrodermal Activity (EDA) measures changes conductivity of the skin which is caused due to sweat gland activity [39]. Since the emotions are closely related to the sympathetic nervous system and increases sweat production in humans. This results in the measurable changes in skin conductivity. Therefore, Skin Conductance Responses (SCRs) and their amplitudes are widely used as an indicator of emotional arousal [7].
3) *Wearables and In-the-Wild Sensing:* The physiological emotion recognition is done by collecting data from humans using compact IoT enabled wearable devices such as smart watches. These devices are equipped with Photoplethysmography (PPG) and Electrodermal Activity (EDA) sensors to monitor emotional states in the real-world scenarios [1].

E. Multimodal Contextual Signals

The human emotions are not dependent on a single data modality but are closely linked to the body movements, gestures and the surrounding environmental contexts [2]

1) *Gestures and Posture:* The gesture and posture features are known as kinesics which include body posture, walking pattern and hand gestures. These provide important information related to emotions particularly in situations where facial expressions are not clear or audio is not available [44]. The gestures and body movements easily show the human emotions. For example, emotions like anger and happiness are generally linked to the fast, energetic and expansive body movements, whereas fear and sadness bring laziness and slower body movements [45]. These patterns and movements are analysed by deep learning-based models which use pose estimation frameworks such as OpenPose. These frameworks extract and study the changes in the body postures or joint angles and their movements to identify the emotional class of an individual [47].
2) *Conversation Context and Environment:* Conversation contextual understanding is another modality used in Emotion recognition. Contextual signals include information related to the conversations such as speaker roles, dialogue history and talking turn patterns. It also includes information related to the lighting conditions, backdrops, scene location and background noise [28]. Inclusion of contextual information helps emotion recognition systems understand the situation of the emotion. This reduces the model hallucinations and improves the prediction accuracy [2].

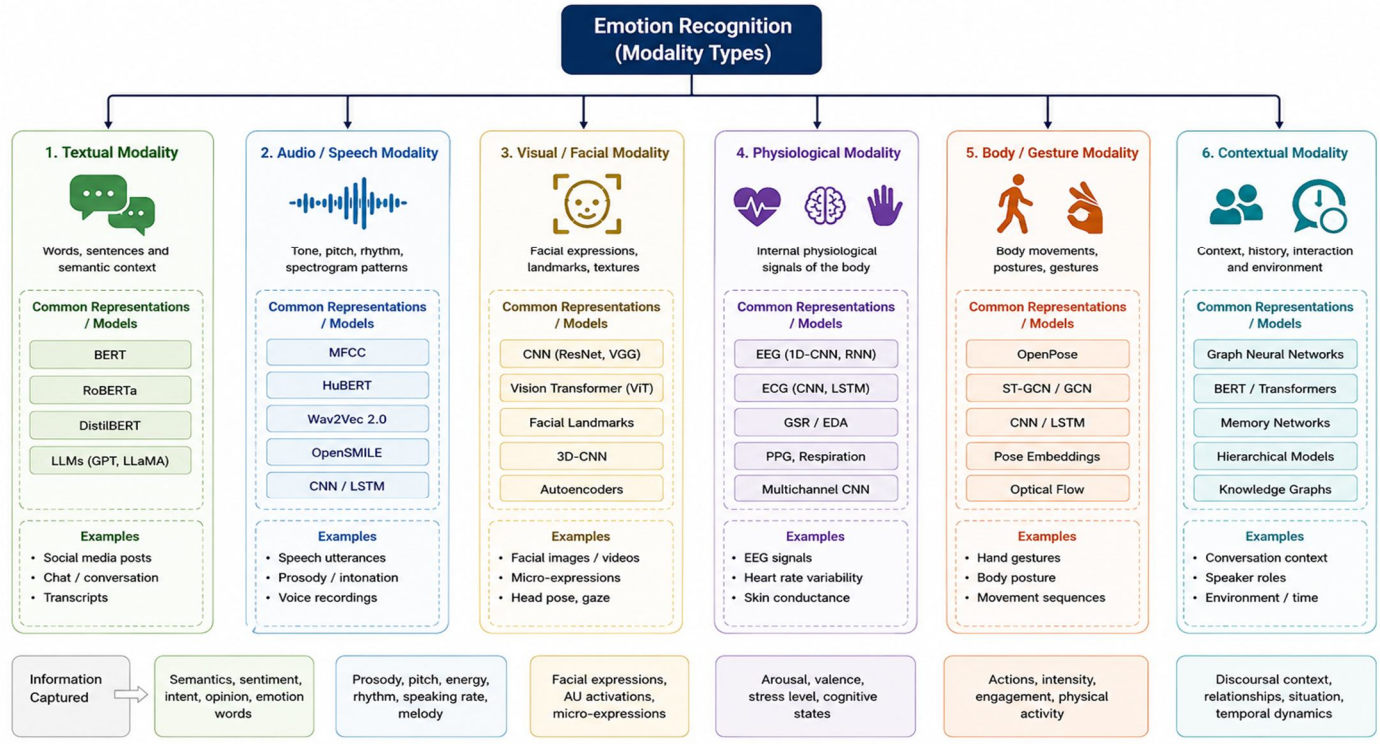


Figure 2: Taxonomy of modalities used in emotion recognition and their models with examples

V. DEEP LEARNING ARCHITECTURES USED IN MER

In the previous section, all the modalities were explained, but these modalities require an appropriate neural network architecture to increase the performance of MER system. Selecting a proper and appropriate architecture influences the ability to process temporal dynamics, handle high-dimensional feature spaces, and execute multi-modal fusion. In this section, the various architecture used in the emotion classification system are discussed.

A. Convolutional and Recurrent Neural Networks (CNNs, RNNs)

Early deep learning approaches for multimodal emotion recognition (MER) simply combined the Convolutional Neural networks (CNN) for spatial feature extraction and Recurrent Neural Networks (RNN) for temporal sequence modelling. The CNN architectures such as ResNet-50 are highly effective in extracting the hierarchical features from images, video frames and audio spectrograms [9]. But these conventional two-dimensional CNNs couldn't capture the temporal relationships between sequential inputs. Therefore, for the purpose of sequence handling where CNNs failed, the RNN models and their advanced variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) networks were widely used in MER systems [1]. These models have hidden memory states that allow them to extract the context-based dependencies between speech sequences and conversations. Despite their effectiveness, LSTM-based architectures face certain limitations such as they often failed to capture very long-term dependencies and require sequential processing of data, which limits the parallel computation of data and increases the overall training time [49].

B. Autoencoders and Generative Adversarial Networks (GANs)

In emotion recognition tasks where, labelled data is limited or signals are highly noisy, deep learning models based on generative architectures are widely used [50]. The various generative architectures used in MER are discussed in this section.

1) *Autoencoders (AEs):* It is one of the generative architectures used and it includes other architectures such as Variational Autoencoders (VAEs) and Vector Quantized VAEs (VQ-VAEs). These architectures are commonly used for unsupervised feature learning in emotion recognition systems. They do so by compressing MER system data in a lower-dimensional space and then generate the original input again. During this time, the network learns the important emotional features automatically and removes noise in the system.
2) *GANs***:** Comprising a competing generator and discriminator, GANs are primarily deployed in MER for data augmentation and modality imputation [50]. In physiological emotion recognition, where subject data (like EEG) is notoriously sparse, GANs can synthesize artificial, high-fidelity physiological waveforms or micro-expressions to balance minority emotion classes and prevent classifier overfitting [52]

C. Graph Neural Networks (GNNs)

GNNs have established themselves as the preeminent architecture for Conversational Emotion Recognition (ERC). In the models like LSTMs, the dialogues are treated as flat sequences, but GNNs model conversations as topological graphs $\boldsymbol{\mathcal{G} = (V, E)}$. Nodes $\boldsymbol{V}$ represent individual multimodal utterances, while edges $\boldsymbol{E}$ define intra-speaker temporal flows and inter-speaker influence interactions. Through a mathematical process known as message passing, node representations are iteratively updated by aggregating the hidden states of adjacent nodes, modified by relation-specific learnable weights $\boldsymbol{W_r}$**:**

$$\boldsymbol{x_i' = W_0 x_i + \sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{1}{|\mathcal{N}_i^r|} W_r x_j}$$

This makes the network easy to track how one speaker's emotional state affects the responses of another speaker For example, an aggressive behaviour of one speaker may trigger a defensive emotional reaction in the other speaker [43]. In addition, the deeper GNN architectures often suffer from the over-smoothing problem where node representations are becoming increasingly similar and lose their ability to distinguish between different utterances [28].

D. Transformers and State Space Models (Mamba)

The Transformer architecture has greatly advanced multimodal emotion recognition due to its multi-head self-attention mechanism. Unlike RNNs, Transformers allow each temporal sequence such as a video frame or a text token to pay attention to all other elements at the same time. This removes the sequential processing limitation of RNNs and enables parallel computation and improves the modelling of long-range contextual relationships [42].

One major limitation of Transformers is the high computational cost of the self-attention mechanism, which requires a time complexity of $\mathcal{O}(N^2)$ operations with respect to the sequence length N. Therefore, in continuous affective computing tasks, where the input may contain thousands of video frames or large-scale EEG recordings, this quadratic complexity quickly exceeds GPU memory capacity [49]. To address this issue, State Space Models (SSMs), particularly the Mamba architecture, has been used recently for the multimodal emotion recognition (MER) [27]. The Mamba architecture retains the information related to contextual modelling capability of Transformers while at the same time it reduces the computational complexity to linear $\mathcal{O}(N)$ scaling through dynamic state-space parameterization [49]. The recent studies from 2024–2025 have shown that Mamba-based MER models are more efficient in processing the continuous multimodal streams than that of Transformer models while still achieving state-of-the-art accuracy [40].

Model	Complexity
Transformer Model	$O(N^2)$
Mamba Model	$O(N)$

E. Multimodal Large Language Models (MLLMs)

One of the most significant developments in the 2025–2026 in the field of multimodal emotion recognition (MER) systems development is the integration of Multimodal Large Language Models (MLLMs) such as GPT-4o and Video-LLaMA. The traditional MER systems that were based on dedicated classification layers to predict emotion categories from a fixed group of emotions but MLLMs process audio, visual, and textual inputs at a same time to generate responses accurately in natural language [15]. This introduces a new concept of emotion reasoning in which model not only identifies emotions but also explains them along with contextual information. Various frameworks have been designed for this purpose such as InstructERC and AffectGPT. These frameworks can perform zero-shot with open-vocabulary inference and produce descriptive outputs like: “The subject appears to be frustrated due to any interrupted workflow which is indicated by increased vocal jitter and furrowed eyebrows” [38].

The development of these frameworks have their own challenges such has fine tuning for large MLLMs is computationally expensive, so the recent MER researches are focusing on efficient parameter adaption techniques. Other models such as Low-Rank Adaption (LoRA) and prompt tuning are used to emotion related knowledge to frozen LLMs [36]. This approach helps in improving the performance without the high cost of training the entire model.

All the models used for emotion detection are summarised in the Table 1.

Table 1: Analysis of different architectures

Architecture	Key Mechanism	Strengths	Weaknesses	Computational Complexity	Suitability
CNN	Spatial convolution & pooling	Excellent spatial feature extraction (images, spectrograms).	Agnostic to temporal dynamics.	Moderate	Frame-level visual/acoustic extraction.
LSTM / GRU	Recurrent hidden state updates	Captures sequential context effectively.	Poor parallelization; vanishing gradients on long sequences.	$\mathcal{O}(N.d^2)$	Short-to-medium utterance modeling.
GNN	Topological message passing	Explicitly models speaker relations and multi-party dialog graphs.	Prone to over-smoothing; memory-intensive for large graphs.	$\mathcal{O}(V + E)$	Conversational Emotion Recognition (ERC).
Transformer	Multi-head self-attention	Global temporal context; excellent parallelization.	Quadratic scaling limits long-video or continuous physiological processing.	$\mathcal{O}(N^2 . d)$	High-resource, complex cross-modal alignment.
Mamba (SSM)	Dynamic state-space parameters	Linear scaling $\mathcal{O}(N)$; matches Transformer accuracy on long sequences.	Complex implementation; nascent adoption in MER.	$\mathcal{O}(N . d)$	Continuous, long-duration affect tracking.
MLLM	Cross-modal adapter + LLM generative decoder	Zero/Few-shot capability; open-vocabulary; provides causal reasoning.	Extreme parameter count; high inference latency; prone to hallucination.	Highly Variable	Explainable AI; Emotion reasoning tasks.

F. Fusion Strategies

The core engineering challenge distinguishing multimodal systems from unimodal baselines is the *fusion strategy*—the specific architectural methodology and mathematical mechanism used to amalgamate disparate sensory inputs [9]. The timing and depth of this interaction heavily dictate the model's resilience to noise.

1) Early, Late, and Hybrid Fusion

a. *Early Fusion (Feature-Level):* Modality-specific features are extracted, temporally synchronized, and immediately concatenated into a single, high-dimensional joint vector prior to the learning stages [9].
   - *Strengths:* Requires training only a single downstream classifier; allows the network to learn fundamental cross-modal correlations from the outset [33].
   - *Weaknesses:* Highly susceptible to the "curse of dimensionality." If modalities exhibit drastically different sampling rates, forced synchronization causes severe information loss or redundancy [9]. Furthermore, it struggles with modality imbalance in which a high-density signal like text can mathematically overpower low-dimensional physiological data.

b. *Late Fusion (Decision-Level):* Each modality is processed through entirely independent neural network branches, each outputting an independent emotion prediction (logits or probabilities) [1]. These independent decisions are subsequently aggregated using rules such as unweighted averaging, majority voting, or adaptive boosting [35].
   - *Strengths:* Highly robust to missing modalities. If the visual sensor fails, the audio and text classifiers still function independently. Excellent for distributed computing architectures.
   - *Weaknesses:* Completely fails to model low-level cross-modal synergies. The network cannot utilize the simultaneous timing of a vocal pitch break and a facial micro-expression because the features never interact.

c. *Hybrid Fusion:* Attempts to capture the benefits of both by initiating feature processing independently, employing interaction layers at intermediate depths, and concluding with a late-fusion weighted decision mechanism [9]. There are several models that follow this strategy. For example, SLSMKCCA (Supervised Least Squares Multiset Kernel Canonical Correlation Analysis) which uses alternating least square optimisation technique to learn the projection matrices. The model then efficiently fuses the facial, acoustic and posture features and also reducing the redundant noise in different modalities [44].

2) Attention-Based and Cross-Modal Transformer Fusion

In order to move forward from combining of static features, the modern MER architectures are increasingly becoming dependent on attention based dynamic fusion mechanisms [2]. This section discusses some of the architectures based on Attention mechanisms.

- *Cross-Modal Attention:* This model is one of the widely used which is based on Transformer architecture. This mechanism learns from weighted relationships between different modalities and dynamically aligns the sematic information which is relevant for emotions. The mechanism of this model works as follows: Consider two asynchronous modalities, such as text ($\boldsymbol{A}$) and video ($\boldsymbol{B}$). The model generates Query matrices ($\boldsymbol{Q_A}$) from the text modality and key ($\boldsymbol{K_B}$) and Value ($\boldsymbol{V_B}$) matrices from the video modality. Their interaction is calculated mathematically using scaled dot-product attention:

$$\textbf{Attention}\left(\mathbf{Q_A}, \mathbf{K_B}, \mathbf{V_B}\right) = \textbf{Softmax}\left(\frac{\mathbf{Q_A}\mathbf{K_B^T}}{\sqrt{\mathbf{d_k}}}\right)\mathbf{V_B}$$

  where $\boldsymbol{d_k}$ represents the scaling factor of key dimension. This framework enables the modality related to linguistic ($\boldsymbol{A}$) to identify the video frames ($\boldsymbol{B}$) which are important for emotion recognition and are linked to the specific spoken words. This shows that fusion is dependent on relevant semantics rather that strict temporal timestamps [32].

3) Contrastive Learning Approaches

The recent studies show that to learn a unified latent representation for multiple modalities, the MER systems are adopting self-supervised and contrastive learning techniques, particularly in the cases where labelled data is limited [14]. Following are the techniques used for this architecture.

- *Cross-Modal InfoNCE:* This is one of the common approaches for Cross-Modal Contrastive Learning (CMCL), which encourages the representations from temporally aligned modalities like synced audio and video to come closed in the latent space and separating unrelated representations. This is typically achieved using the InfoNCE loss function:

$$\mathcal{L}_{\text{NCE}}^{\mathcal{A}\to\mathcal{B}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x_i, y_j)/\tau)}$$

where $\text{sim}(\cdot)$ function denotes cosine similarity, (x_i, y_i) represents a positive pair of aligned multimodal samples and τ is the temperature parameter controlling the sharpness of the similarity distribution. The loss is reduced by maximizing similarity between related samples and minimizing similarity between unrelated ones, CMCL helps the model learn more robust and semantically meaningful multimodal representations [25, 26].

- *Supervised Contrastive Learning (SupCon):* SupCon learning framework is used when the emotion labels are available. It works by extending the contrastive learning by grouping the samples according to their emotion category rather than their modality. In this approach, samples which are sharing the same emotional label are encouraged to cluster closely in the latent space, but the samples from different emotion classes are pushed farther away. This improves emotion class separation and representation in data modalities.

$$\mathcal{L}_{\text{Supcon}} = \sum_{i=1}^{N}\frac{-1}{|P(i)|}\sum_{p\in P(i)} log\frac{\exp(z_i \cdot z_p/\tau)}{\sum_{a\in A(i)}\exp(z_i \cdot z_a/\tau)}$$

where $P(i)$ denotes the set of positive samples sharing the same emotion label as sample i, $A(i)$ represents all candidate samples of a particular batch, z_i is the learned representation of sample i, and τ is the temperature parameter. By aligning samples according to their emotional semantics rather than modality type, **SupCon** enables the model to learn more discriminative and robust multimodal representations, leading to improved downstream emotion classification performance [16,17]

G. *Benchmark Datasets*

A large, diverse and well-labelled dataset is necessary for any of the AI or machine learning model for its accuracy. In case of MER system following are the benchmark datasets that are being used for training and evaluating the systems for emotion prediction. [21]

Table 2: Analysis of different datasets

Dataset Name	Modalities	Samples / Size	Annotation Type	Application Area	Limitations
IEMOCAP [1]	Audio, Video, Text, MoCap	10,039 utterances	Discrete (Happy, Angry, Sad, Neutral)	Conversational Analysis	Highly scripted/acted; limited demographic variance (10 actors); dyadic only.
MELD [54]	Audio, Video, Text	~13,000 utterances	Discrete (7 emotions) & Sentiment	Multi-party Conversational	Sourced from a TV sitcom ("Friends"); lacks genuine spontaneous affect.
CMU-MOSEI [1]	Audio, Video, Text	23,500 segments	Continuous (V-A) & Discrete	Sentiment & Emotion Inference	Biased toward the demographics of YouTube vloggers; highly variable acoustic noise.
RAVDESS [55]	Audio, Video	2,452 vocal clips	Discrete (8 emotions, 2 intensities)	Speech & Song Emotion	Strictly controlled laboratory acoustics; emotions are heavily exaggerated.
DEAP [2]	EEG, ECG, GSR, Video	32 subjects (40 trials each)	Continuous (Valence, Arousal, Dominance)	Physiological Affect	Small participant pool limits cross-subject generalization capabilities.
SEED [21]	EEG, Eye-tracking	15 subjects	Discrete (Positive, Negative, Neutral)	Neuro-cognitive ER	Extremely limited discrete emotional classes; culturally homogeneous subjects.
MAHNOB-HCI [22]	EEG, Physio, Eye, Video	27 subjects	Discrete (9 emotions) & V-A	Human-Computer Interaction	Data recorded under highly constrained, unnatural lab conditions.

Dataset Biases and Annotation Inconsistencies

A critical analysis of the existing benchmark datasets discussed in *Table 2* shows various important limitations [21]. Most these datasets were dominated by participants from WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations [31]. For example, the models trained on datasets such as CMU-MOSEI showed reduced performance when tested on non-Western populations [14].

Another major challenge is data annotation inconsistency and this effects the reliability of truth labels. Emotion annotation is largely dependent on subjective interpretation of human annotators. The annotation label is usually measured via Inter-Rater Reliability measures such as Cohen's κ. The annotators usually agree on the strongly expressed emotions, such as extreme joy or anger, but the

emotions with uncertain boundary like frustration versus anger or amusement versus happiness, most of time leads to disagreements and are labelled as noise [51] This inconsistency negatively effects the training and optimisation of MER models.

H. Evaluation Metrics

The evaluation of multimodal emotion recognition (MER) systems is challenging due to the complex nature of emotional data. Emotional datasets often contain class imbalance and continuously changing emotional states [9]. Therefore, traditional machine learning evaluation metrics alone are not sufficient for comprehensive performance assessment.

1) Classification Metrics

a. *Accuracy:* It is defined as the ration of correctly predicted samples to the total number of samples. It is one of the most commonly used evaluation method but it can be highly misleading and deceptive due to class imbalance in datasets. For example, if "Neutral" category represents a large portion of dataset and other emotions such as "Fear" may occur rarely. The model trained over such a dataset will perform with high accuracy on the "Neutral" class as it is majority class, but performs poorly with the other actual emotions. Therefore, accuracy alone is not sufficient for evaluating modern MER system and is not single handedly used to measure the performance [9].

b. *Precision, Recall, and F1-score:*
The problem of imbalance in classification metrics is handled by the use of Precision, recall and F1-score as evaluation metrics. Among all of these the F1-score is widely preferred because it creates a proper balance between both false positives and false negatives. F1-score is simply the harmonic mean of precision and recall. More specifically, the *Macro-F1* is calculated by the average of F1-score of all the emotion classes by giving equal priority to each class. This is useful as it evaluates how model performs on emotional classes which are in minority. The other F1 known as *Weighted-F1* assigns weights according to the number of samples of each class, which creates a more balanced measure of overall system performance. The weighted-F1 is very much effective in handling the imbalanced datasets, therefore it is commonly used as a primary evaluation metric in datasets such as IEMOCAP and MELD [9].

c. *ROC-AUC (Receiver Operating Characteristic - Area Under Curve):* This method is used for binary sentiment classification and multi-label emotion recognition tasks. ROC-AUC measures the probability that a classifier ranks a randomly selected positive sample higher than a randomly selected negative sample. Since it evaluates performance across different thresholds, it provides a threshold-independent measure of the model's discriminative ability.

2) Regression Metrics for Dimensional Affect

The datasets which are based on Valence-Arousal representations such as DEAP and SEMAINE, the traditional metrics like Mean Squared Error (MSE) are insufficient for evaluating emotional dynamics. The MSE measures point-wise prediction error, but it fails to capture the temporal variation of emotional responses with time. For example, a model which predicts the average emotional value may still a low MSE due to missingness of important fluctuations and transitions of the emotion signal [23].

- *Concordance Correlation Coefficient (CCC)***:** CCC is widely used for dimensional affect regression tasks (prediction tasks where emotions are represented as continuous values rather than fixed categories) because it measures the agreement between the predicted sequence and the actual ground-truth sequence [9]. Unlike the other standard correlation metrics, the CCC framework considers both correlation and differences in mean and scale between predictions and actual values. It is defined as:

$$\boldsymbol{\rho_c} = \frac{\boldsymbol{2\sigma_{xy}}}{\boldsymbol{(\mu_x - \mu_y)^2 + \sigma_x^2 + \sigma_y^2}}$$

where $\boldsymbol{\mu}$ represents the mean values, $\boldsymbol{\sigma^2}$ denotes the variances, and $\boldsymbol{\sigma_{xy}}$ represents the covariance between the predicted and the ground truth sequences. A higher CCC value means better agreement between predicted and actual emotional trajectories [20].

I. Current Challenges

The MER systems have achieved a remarkable result on controlled benchmark datasets. But their prediction power often decreases in real world scenarios, because of incomplete, unpredictable or noisy data [1]. There are various challenges which continue to limit the practical working of MER systems.

1) *Modality Imbalance and Missing Modalities:* In real world deployment of MER systems, data from different modalities is not always available or reliable. For example, user may go out of the camera zone or noisy backgrounds distorts audio signals, or the wearable sensors may malfunction or lose contact with the human body. These issues create missing data or corrupted data modalities, which effect the overall performance of the traditional early fusion models that were dependent of complete input data . This problem is addressed by a solution known as dynamic imputation techniques in which science and statistics is used to fill the missing values in dataset. The frameworks like Dynamic Modality Recognition and Enhancement (DREAM) and Hyper-Modality Enhancement (HME) attempt to reconstruct the missing modality using diffusion models or cross model information before the fusion stage . The other issue is modality imbalance in which the model relies too mostly on the dominant modality. This limits the model's ability to capture the emotion information from all the modalities effectively [13].
2) *Noisy Data and Cross-Cultural Understanding:* The real-world datasets collected in uncontrolled environments (like changing camera angles, background noise etc) make emotion recognition very difficult [31]. More importantly emotion expressing varies across different cultures. Therefore, the models trained on Western datasets may fail to interpret the emotions correctly. [14]
3) *Privacy, Ethical Concerns, and Explainability (XAI):* The MER systems collect sensitive information such as facial expressions, EEG signals, ECG signals etc, which raises a serious privacy and ethical concerns [29]. This information reveals personal emotional and mental state, therefore strict regulations for data usage and transparency are required. Another problem with the deep learning models, Transformer models and MLLMs is that they work as "black boxes", which means that their decision-making process is very difficult to understand. This puts a limitation to their use in highly sensitive areas such as healthcare, legal matters and security where explanations are essentially required for ever decision made by the system [19]. The solution for transparency is explained to some extent by Explainable AI (XAI) techniques which include Grad-CAM and Layer-wise Relevance Propagation (LRP) are used to highlight which part of the input data is used by the model for the emotion prediction [2]. However, fully connecting these technical explanations with human psychological reasoning still remains a significant challenge.
4) *Real-Time Deployment and Low-Resource Settings:* This problem is one of the major challenges which limits the use of AI in real world decision making on the basis of emotional activity of an individual. The large Transformer models and MLLMs require too high computational power and memory, and makes it difficult to be used on small wearable devices which work on less power and memory [1]. Due to small memory, it becomes very difficult for such devices to work on high dimension modalities during emotion inference [18]. The techniques like model quantisation, pruning and knowledge distillation are used to reduce the model size and work properly with less power and memory.

J. Emerging Trends and Future Directions

The future work of MER is to develop models that can are understandable, decentralised and can be easily integrated into real world environments.

1) *Emotion-Aware Multimodal Foundation Models (MLLMs) and Emotion Reasoning:* The most important recent trend in MER systems is shift from traditional fixed set classifications to the generalised MLLMs [15]. The frameworks used for this purpose are AffectGPT, InstructERC, and EmoLLM understand and explain the emotions in human-like manner [38]. Instead of only predicting an emotion label, these models can explain the reason behind the emotion by connecting the various modalities of the emotion such as facial expressions, body gestures, voice tempo changes etc [15]. Future research will focus heavily on parameter-efficient tuning (e.g., multimodal LoRA) to inject emotion-specific competencies into these massive foundation models without catastrophic computational costs [14].
2) *Federated Learning (FL) and Edge AI:* To address the need for large and diverse real-world emotional datasets while still protecting sensitive biometric information, Federated Learning is emerging as an important solution in MER systems [17]. In this method, the sensitive personal data is kept on the user's local edge device and model is trained over that device, only the learned model updates or gradients are shared with a central system for optimisation [11]. Current research is also focusing improving these aggregation methods. For example, FedAvg performs well when the training data across devices is similar and other algorithms such as FedProx are designed to handle the highly diverse and non-identically distributed nature of human emotional data more effectively [6].

3) Affective Conversational Agents, Embodied AI, and Social Robotics: The main goal of the affective computing is developing an Embodied AI, where the system predicts the human emotions and then physically interacts with human environments. In social robotics, affective computing helps create a complete interaction loop between humans and machines. For example, a robot equipped with multimodal sensors can detect an elderly patient's frustration, understand the possible reason behind it, and adjust its behaviour accordingly by changing its distance, tone of voice, or movement speed in order to reduce stress and improve trust [7].

Achieving smooth, real-time, and empathetic interaction between humans and intelligent physical systems is considered one of the most important future directions of human centered artificial intelligence [8].

K. Comparative Analysis Section

1) Timeline of Evolution

The development of Multimodal Emotion Recognition (MER) can be clearly divided into four major technological eras [2]:

a. Pre-2015 (Classical Era): This period was dominated by unimodal emotion recognition research where models were dependent on handcrafted LLDs (MFCCs, LBP, FACS) and shallow SVMs. Late fusion was standard due to hardware limitations [1].

*b. 2015-2019 (Deep Learning Integration)***:** The widespread adoption of CNNs for spatial extraction and LSTMs/RNNs for temporal sequencing. End-to-end learning replaces hand-engineered features [2]. Graph Neural Networks emerge specifically to solve Conversational Emotion Recognition (ERC) topologies.

c. 2020-2023 (The Transformer Era): Self-supervised unimodal encoders (Wav2Vec2, RoBERTa, ViT) drastically elevate baseline extraction. Cross-modal attention mechanisms mathematically resolve asynchronous fusion issues, leading to widespread early/hybrid fusion architectures [2].

d. 2024-2026 (Foundation Models & Reasoning): The rise of generative MLLMs, State-Space Models (Mamba) for linear-time continuous sensing, and self-supervised contrastive learning. The field shifts decisively from closed-set classification to open-vocabulary, causally driven emotion reasoning [49].

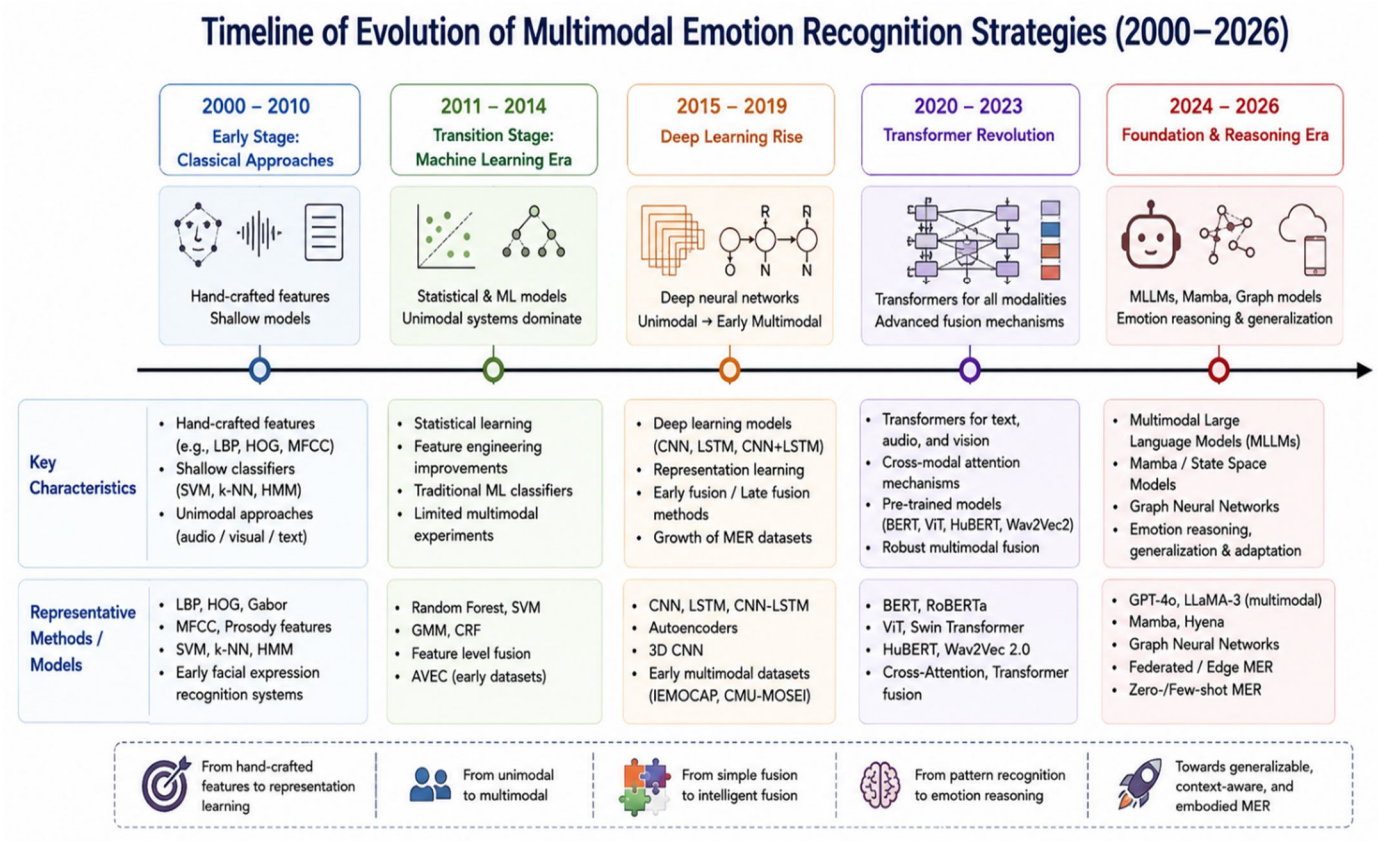


Figure 3: Timeline of evolution of MER strategies from 2000 to 2026

2) Performance Comparison of SOTA Architectures

Table 3 provides a comparative overview of state-of-the-art models evaluated on the preeminent conversational benchmarks, IEMOCAP and MELD.

Table 3: Comparison of different MER models used on IEMOCAP and MELD datasets

Model / Architecture	Fusion Strategy	Modalities	IEMOCAP (W-F1)	MELD (W-F1)	Publication / Source
DialogueGCN	GNN / Contextual Message Passing	A, V, T	~64.20%	~58.10%	Baseline Standard [46]
EmoBERTa-CNN	RoBERTa + CNN / Hybrid	T, V	N/A	79.45%	[48]
InstructERC	MLLM / Retrieval-Augmented Seq2Seq	T (Proxy)	71.39%	-	[38]
GraphSmile	GNN + Sentiment Dynamics	A, V, T	72.81%	-	[53]
Mamba-like Models	State-Space Model + Probabilistic Fusion	A, V, T	73.30%	-	[27]
GS-MCC	GNN + Fourier Contrastive Learning	A, V, T	73.90%	-	[14]

Abbreviation Key: A = Audio; V = Video; T = Text. T (Proxy) = acoustic or visual features were transformed into textual descriptions

L. Conclusion

Multimodal Emotion Recognition represents a rapidly maturing intersection of artificial intelligence, cognitive neuroscience, and human-centered design. The field has shown transition from classical statistical and machine learning approaches to advanced deep learning architectures such as Graph Neural Networks (GNNs), Transformers, State Space Models like Mamba, and Multimodal Large Language Models (MLLMs). These models have improved the ability of computers to understand and interpret complex emotions. Modern models which are based on fusion techniques particularly cross-attention mechanisms and self-supervised contrastive learning, have removed major challenges like feature redundancy and semantic misalignment between different modalities. Despite these advancements, MER systems still face too many challenges when the models are moved from laboratory to real-world scenarios. Also, most of the datasets are culturally biased or scripted and are collected under constrained conditions. The modern researches are trying to address these problems along with other problems such as handling missing modalities, reducing cross cultural bias and optimisation of computationally expensive MLLMs for real time Edge AI applications without reducing the accuracy.

As MER systems are becoming increasingly important for integration into daily life in the fields such as mental health monitoring, healthcare, virtual assistants and social robotics. The issues related to privacy, fairness and transparency are becoming critically important. The Federated Learning are helping to protect the sensitive user data while Explainable AI (XAI) methods are essential for maintaining trust and interpretability in the automatic emotion recognition systems. By addressing all these challenges, MER has the potential to contribute significantly toward the development of intelligent, empathetic, and socially aware artificial systems capable of interacting naturally and effectively with humans.

WORKS CITED

[1] M. J. D. Kumar, M. Sukesh Rao, and K. C. Narendra, "Multimodal Emotion Recognition: A Comprehensive Survey of Datasets, Methods, and Applications," IEEE Access, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3636186.

[2] Wu Y, Mi Q, Gao T. A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. Biomimetics (Basel). 2025 Jun 27;10(7):418. doi: 10.3390/biomimetics10070418. PMID: 40710231; PMCID: PMC12292624.

[3] 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 (Basel). 2023 Oct 12;25(10):1440. doi: 10.3390/e25101440. PMID: 37895561; PMCID: PMC10606253.

[4] Ding S, Ma L and Li H (2025) Multimodal physiological signal emotion recognition based on multi-head cross attention with representation learning. Front. Psychiatry 16:1713559. doi: 10.3389/fpsyt.2025.1713559

[5] Yu, L.; Ge, Y.; Ansari, S.; Imran, M.; Ahmad, W. Multimodal Sensing-Enabled Large Language Models for Automated Emotional Regulation: A Review of Current Technologies, Opportunities, and Challenges. Sensors 2025, 25, 4763. https://doi.org/10.3390/s25154763

[6] Ma, X., “Comparative Analysis of FedAvg and FedProx Algorithms in Federated Learning for Handwritten Character Recognition on the EMNIST Dataset”. Academic Journal of Science and Technology, 19(2), 501-506. https://doi.org/10.54097/h7srvr13

[7] F. Rahimi, C. Tamantini, A. Orlandini, F. Fracasso, and R. Siciliano, "Comparing Fusion Strategies for Multimodal Emotion Prediction Using Deep Physiological Features," in Proc. Workshop on Social Robotics for Human-Centered Assistive and Rehabilitation AI (Fit4MedRob), held in conjunction with the International Conference on Social Robotics (ICSR), 2025.

[8] Emily S. Cross, Arvid Kappas. 2026. Social Robotics Is Not (Just) About Machines, It Is About People: Psychology's Role in Developing Social Machines. Annual Review Psychology. 77:649-678. https://doi.org/10.1146/annurev-psych-040325-025951

[9] A.-S. Moon, H. Kim, Y.-C. Park, and J. Lee, "A Survey on Multimodal Emotion Recognition: Methods, Datasets, and Future Directions," Computers, Materials & Continua, vol. 87, no. 2, 2026, doi: 10.32604/cmc.2026.076411.

[10] Y. Shou, T. Meng, W. Ai, F. Fu, N. Yin, and K. Li, "A Comprehensive Survey on Multi-modal Conversational Emotion Recognition with Deep Learning," arXiv preprint
arXiv:2312.05735, 2025, doi: 10.48550/arXiv.2312.05735.

[11] A. Nandi and F. Xhafa, "A federated learning method for real-time emotion state classification from multi-modal streaming," Methods, vol. 204, pp. 340–347, Aug. 2022, doi: 10.1016/j.ymeth.2022.03.005.

[12] A. Yazici, T. Kucukyilmaz, T. Dokeroglu, A. Sharipbay, M. H. Lee, and B. Tyler, “State-of-the-art Multimodal Emotion Recognition: A comprehensive survey and taxonomy,” Intelligent Systems with Applications, vol. 30, Art. no. 200642, 2026, doi: 10.1016/j.iswa.2026.200642.

[13] Lanxin Bi, Yunqi Zhang, Luyi Wang, Yake Niu, and Hui Zhao. 2025. Two Challenges, One Solution: Robust Multimodal Learning through Dynamic Modality Recognition and Enhancement. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12855–12867, Suzhou, China. Association for Computational Linguistics.

[14] Chengyan Wu et al, Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects - ACL

[15] J. Han et al., "Pioneering Multimodal Emotion Recognition in the Era of Large Models: From Closed Sets to Open Vocabularies," arXiv preprint arXiv:2512.20938, 2025, doi: 10.48550/arXiv.2512.20938.

[16] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. In: Advances in Neural Information Processing Systems. 2020; 33:18661–18673.

[17] Che L, Wang J, Zhou Y, Ma F. Multimodal Federated Learning: A Survey. Sensors (Basel). 2023 Aug 6;23(15):6986. doi: 10.3390/s23156986. PMID: 37571768; PMCID: PMC10422520.

[18] Mostert, W.; Kurien, A.; Djouani, K. Multi-Modal Emotion Detection and Tracking System Using AI Techniques. Computers 2025, 14, 441. https://doi.org/10.3390/computers14100441

[19] S. Sarah et al., "Multimodal Emotion Recognition with Explainable AI for Cognitive Human-Computer Interaction in Smart Environments," 2025 5th International Conference on Soft Computing for Security Applications (ICSCSA), Salem, India, 2025, pp. 1091-1096, doi: 10.1109/ICSCSA66339.2025.11170860

[20] Lin, L. I.. “A concordance correlation coefficient to evaluate reproducibility. *Biometrics”*, 45(1), 255-268

[21] Grosu, M.-M.; Datcu, O.; Tapu, R.; Mocanu, B. A Comparative Study of Emotion Recognition Systems: From Classical Approaches to Multimodal Large Language Models. Appl. Sci. 2026, 16, 1289. https://doi.org/10.3390/app16031289

[22] A. Hoffsommer, H. Schneider, S. Pavlitska, and J. M. Zöllner, "DEAP DIVE: Dataset Investigation with Vision Transformers for EEG Evaluation," arXiv preprint arXiv:2510.00725, 2025, doi: 10.48550/arXiv.2510.00725.

[23] B. T. Atmaja and M. Akagi, "Evaluation of Error and Correlation-Based Loss Functions for Multitask Learning Dimensional Speech Emotion Recognition," arXiv preprint arXiv:2003.10724, 2020, doi: 10.48550/arXiv.2003.10724

[24] Mengara Mengara, A.G.; Moon, Y.-k. CAG-MoE: Multimodal Emotion Recognition with Cross-Attention Gated Mixture of Experts. Mathematics 2025, 13, 1907. https://doi.org/10.3390/math13121907

[25] H. Zhang et al., "Cross-Modal Contrastive Learning for Text-to-Image Generation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 833–842.

[26] Y. Zhao et al., "Heterogeneous Interactive Graph Network for Audio–Visual Question Answering," Knowledge-Based Systems, vol. 300, Art. no. 112165, 2024.

[27] Shou, Y., Meng, T., Ai, W., Li, K. (2026). Revisiting Multi-modal Emotion Learning with Broad State Space Models and Probability-Guidance Fusion. In: Ribeiro, R.P., et al. Machine Learning and Knowledge Discovery in Databases. Research Track. ECML PKDD 2025. Lecture Notes in Computer Science(), vol 16016. Springer, Cham. https://doi.org/10.1007/978-3-032-06078-5_29

[28] Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. 2025. Dynamic Graph Neural ODE Network for Multi-modal Emotion Recognition in Conversation. In Proceedings of the 31st International Conference on Computational Linguistics, pages 256–268, Abu Dhabi, UAE. Association for Computational Linguistics.

[29] Barker, D.; Tippireddy, M.K.R.; Farhan, A.; Ahmed, B. Ethical Considerations in Emotion Recognition Research. Psychol. Int. 2025, 7, 43. https://doi.org/10.3390/psycholint7020043

[30] XUE Jieying, Emotion Detection with Context, Emotional Dynamics, and Speaker Personality Modeling, JAIST Repository [Online]

[31] S. Kalateh, L. A. Estrada-Jimenez, S. Nikghadam-Hojjati, and J. Barata, "A Systematic Review on Multimodal Emotion Recognition: Building Blocks, Current State, Applications, and Challenges," IEEE Access, vol. 12, pp. 104000–104035, 2024, doi: 10.1109/ACCESS.2024.3430850

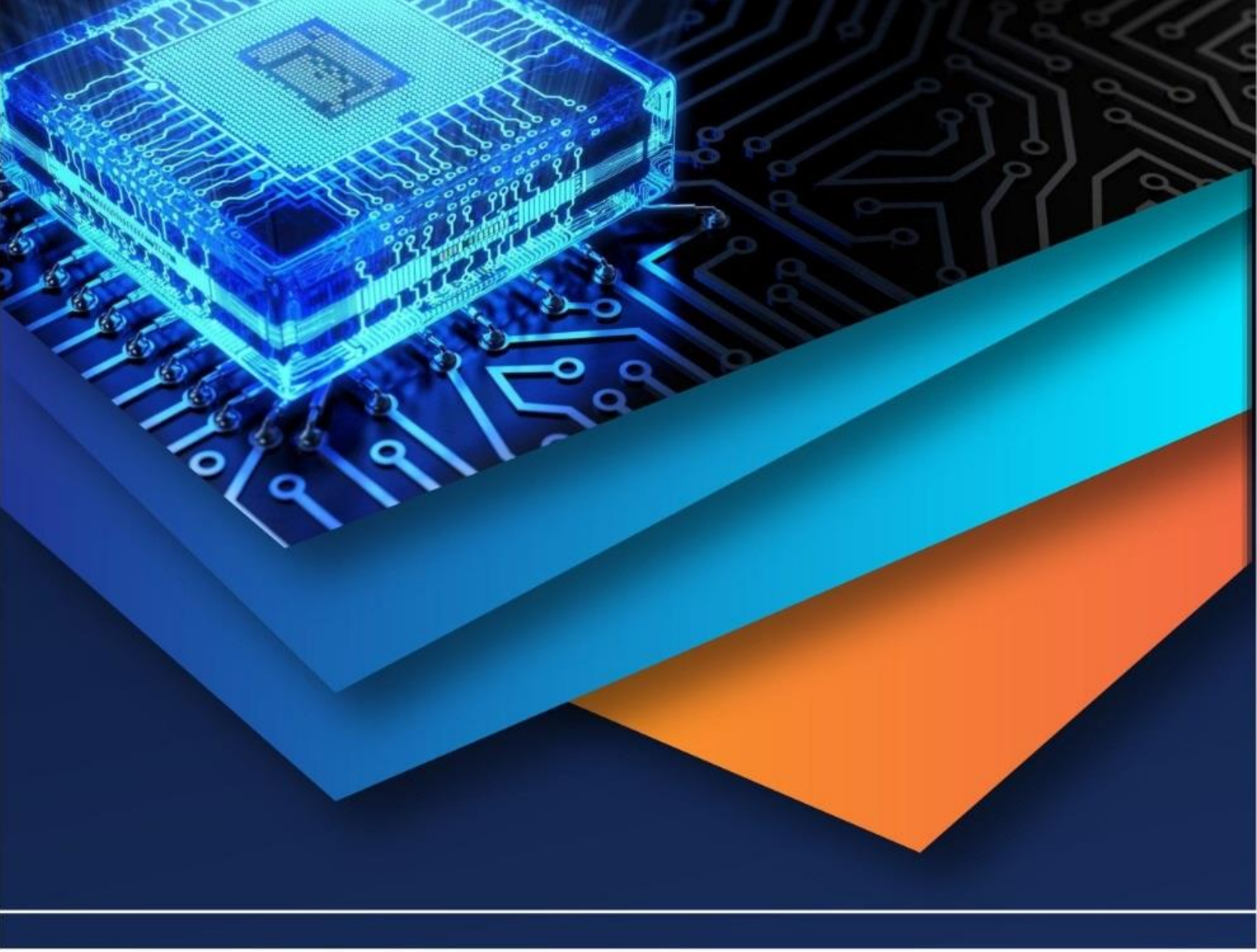
[32] G. Seneviratne et al., "CROSS-GAiT: Cross-Attention-Based Multimodal Representation Fusion for Parametric Gait Adaptation in Complex Terrains," arXiv preprint arXiv:2409.17262, 2024.

[33] Moorthy, S.; Moon, Y.-K. Hybrid Multi-Attention Network for Audio–Visual Emotion Recognition Through Multimodal Feature Fusion. Mathematics 2025, 13, 1100. https://doi.org/10.3390/math13071100

[34] R. Zhao et al., "Leveraging Cross-Attention Transformer and Multi-Feature Fusion for Cross-Linguistic Speech Emotion Recognition," arXiv preprint arXiv:2501.10408, 2025.

[35] S. R. Ahamed et al., "Evaluating Early, Late and Hybrid Fusion in Multimodal Emotion Detection with Pretrained Models," Research Square, Apr. 2026, doi: 10.21203/rs.3.rs-8907947/v1

[36] Y. Sun and T. Zhou, "DialogueMLLM: Transforming Multimodal Emotion Recognition in Conversation Through Instruction-Tuned MLLM," IEEE Access, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3591447.
[37] Shuai, T.; Beng, S.; Khalid, F.B.; Rahmat, R.W.B.O.K. Advances in Facial Micro-Expression Detection and Recognition: A Comprehensive Review. Information 2025, 16, 876. https://doi.org/10.3390/info16100876
[38] S. Lei et al., "InstructERC: Reforming Emotion Recognition in Conversation with Multi-task Retrieval-Augmented Large Language Models," arXiv preprint arXiv:2309.11911, 2024.
[39] D. M. L. Dissanayake, "Emotion Recognition from Physiological Signals Using Machine Learning on the CASE Dataset," M.S. thesis, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland, Dec. 2025. [Online]
[40] J. Murzaku and O. Rambow, "OmniVox: Zero-Shot Emotion Recognition with Omni-LLMs," arXiv preprint arXiv:2503.21480, Mar. 2025. Available: https://doi.org/10.48550/arXiv.2503.21480
[41] Karthiga M, Suganya E, Sountharrajan S, Balusamy B, Selvarajan S. Eeg based smart emotion recognition using meta heuristic optimization and hybrid deep learning techniques. Sci Rep. 2024 Dec 4;14(1):30251. doi: 10.1038/s41598-024-80448-5. PMID: 39632923; PMCID: PMC11618626.
[42] B. L. Fuchs et al., "Understanding Transformer Reasoning Capabilities via Graph Algorithms," in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, 2024.
[43] Liu J, Li J, Dong J, Mo Z, Liu N, Li Q, Yuan Y. Adaptive Graph Learning with Multimodal Fusion for Emotion Recognition in Conversation. Biomimetics (Basel). 2025 Jun 25;10(7):414. doi: 10.3390/biomimetics10070414.
[44] Yan, J.; Li, P.; Du, C.; Zhu, K.; Zhou, X.; Liu, Y.; Wei, J. Multimodal Emotion Recognition Based on Facial Expressions, Speech, and Body Gestures. Electronics 2024, 13, 3756. https://doi.org/10.3390/electronics13183756
[45] Kipp, M., & Martin, J. C. (2015). Expressing emotion through posture and gesture. In R. A. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.), The Oxford Handbook of Affective Computing (pp. 209–221). Oxford University Press.
[46] Xie, J.; Wang, Y.; Meng, T.; Tai, J.; Zheng, Y.; Varatnitski, Y.I. Multimodal Emotion Recognition Method Based on Domain Generalization and Graph Neural Networks. Electronics 2025, 14, 885. https://doi.org/10.3390/electronics14050885
[47] H. Liu, "Emotion Detection through Body Gesture and Face," arXiv preprint arXiv:2407.09913, 2024, doi: 10.48550/arXiv.2407.09913.
[48] Zhang, M.; Yu, A.; Sheng, X.; Park, J.; Rhee, J.; Cho, K. EmoBERTa–CNN: Hybrid Deep Learning Approach Capturing Global Semantics and Local Features for Enhanced Emotion Recognition in Conversational Settings. Mathematics 2025, 13, 2438. https://doi.org/10.3390/math13152438
[49] A. Koledoye, C. Unachukwu, G. Nwobu, and H. Rana, "Benchmarking the Computational and Representational Efficiency of State Space Models against Transformers on Long-ContextDyadic Sessions," arXiv preprint arXiv:2601.01237, 2026, doi: 10.48550/arXiv.2601.01237.
[50] F. Ma et al., "A Review of Human Emotion Synthesis Based on Generative Technology" in IEEE Transactions on Affective Computing, vol. 16, no. 04, pp. 2579-2598, Oct.-Dec. 2025, doi: 10.1109/TAFFC.2025.3573878.
[51] Y. Shou, T. Meng, W. Ai, and K. Li, "Multimodal Large Language Models Meet Multimodal Emotion Recognition and Reasoning: A Survey," arXiv preprint arXiv:2509.24322, 2025, doi: 10.48550/arXiv.2509.24322.
[52] Hegh, A. N., Adeyelu, A. A., Iorliam, A., & Otor, S. U. (2025). MULTI-MODAL EMOTION RECOGNITION MODEL USING GENERATIVE ADVERSARIAL NETWORKS (GANs) FOR AUGMENTING FACIAL EXPRESSIONS AND PHYSIOLOGICAL SIGNALS. FUDMA JOURNAL OF SCIENCES, 9(5), 277-290. https://doi.org/10.33003/fjs-2025-0905-3412
[53] J. Li, X. Wang and Z. Zeng, "Tracing Intricate Cues in Dialogue: Joint Graph Structure and Sentiment Dynamics for Multimodal Emotion Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, no. 10, pp. 8786-8803, Oct. 2025, doi: 10.1109/TPAMI.2025.3581236
[54] Soujanya Poria et al. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations" Doi: https://doi.org/10.48550/arXiv.1810.02508
[55] Steven R. Livingstone, Frank 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" doi: https://doi.org/10.1371/journal.pone.0196391



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)